RAPID|MINER

BENUTZERHANDBUCH

rapid-i
REPORT THE FUTURE

RapidMiner 5.0

Benutzerhandbuch

Rapid-I
www.rapid-i.com

Dieses Werk ist urheberreichtlich geschützt. Alle Rechte, auch die der Übersetzung, des Nachdrucks und der Vervielfältigung des Buches, oder Teilen daraus, vorbehalten. Kein Teil des Werkes darf ohne schriftliche Genehmigung in irgendeiner Form reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Copyright ©2010 Rapid-I

Inhaltsverzeichnis

1 Grundbegriffe **1**

1.1 Zufall oder nicht? 1

1.2 Grundbegriffe 5

1.2.1 Attribute und Zielattribute 6

1.2.2 Konzepte und Beispiele 9

1.2.3 Attributrollen 11

1.2.4 Wertetypen 11

1.2.5 Daten und Metadaten 14

1.2.6 Modellierung 15

2 Design **19**

2.0.1 Flexibilität und Funktionsvielfalt 20

2.0.2 Skalierbarkeit 21

2.0.3 Eine Frage des Formats 22

2.1 Installation und Erstes Repository 22

2.2 Perspektiven und Views 24

2.3 Design-Perspektive 29

2.3.1 Operators und Repositories View 30

2.3.2 Process View 34

2.3.3 Operatoren und Prozesse 35

2.3.4 Weitere Optionen des Process Views 44

2.3.5 Parameters View 46

2.3.6 Help und Comment View 49

2.3.7 Overview View 51

2.3.8 Problems und Log View 52

3 Analyseprozesse **57**

3.1 Erstellen eines neuen Prozesses 57

3.2 Der erste Analyseprozess 60
3.2.1 Transformation der Metadaten 63
3.3 Ausführung von Prozessen 72
3.3.1 Betrachten von Ergebnissen 74
3.3.2 Breakpoints 76

4 Darstellung **79**
4.1 Systemmonitor 79
4.2 Anzeigen von Ergebnissen 81
4.2.1 Quellen für die Anzeige von Ergebnissen 82
4.3 Über Datenkopien und Views 84
4.4 Darstellungsformen 85
4.4.1 Text 86
4.4.2 Tabellen 86
4.4.3 Plotter 92
4.4.4 Graphen 94
4.4.5 Spezielle Ansichten 97
4.5 Result Overview 97

5 Repository **99**
5.1 Das RapidMiner Repository 99
5.1.1 Ein neues Repository anlegen 101
5.2 Das Repository verwenden 102
5.2.1 Prozesse und relative Repositoryangaben 103
5.2.2 Daten und Objekte in das Repository importieren 104
5.2.3 Zugriff und Verwaltung des Repositories 107
5.2.4 Der Prozesskontext 108
5.3 Daten und Metadaten 109
5.3.1 Metadatenpropagierung vom Repository durch den Prozess 112

1 Motivation und Grundbegriffe

In diesem Kapitel möchten wir Ihnen eine kleine Motivation für den Einsatz von Data Mining an die Hand geben und ganz nebenbei auch noch die wichtigsten Begriffe einführen. Ob Sie nun bereits erfahrener Data Mining Experte sind oder nicht – die Lektüre dieses Kapitels lohnt sich in jedem Fall, damit Sie die sowohl hier als auch die in RapidMiner verwendeten Terme kennen und beherrschen.

1.1 Zufall oder nicht?

Bevor wir nun richtig starten, versuchen wir noch ein kleines Experiment:

- Denken Sie sich eine Zahl zwischen 1 und 10.
- Multiplizieren Sie diese Zahl mit 9.
- Bilden Sie die Quersumme des Ergebnisses, also die Summe der Ziffern.
- Multiplizieren Sie das Ergebnis mit 4.
- Teilen Sie das Resultat durch 3.
- Ziehen Sie 10 ab.

Das Ergebnis ist 2.

Glauben Sie an Zufall? Als Analyst werden Sie diese Frage wohl verneinen lernen oder tun dies sogar bereits. Nehmen wir beispielsweise das wohl einfachste Zufallsereignis, dass man sich überhaupt nur vorstellen kann, nämlich den Wurf

einer Münze. „Aha“ mögen Sie denken, „das ist doch ein zufälliges Ereignis und niemand kann vorhersagen, welche Seite der Münze nach einem Wurf oben liegt“. Das stimmt zwar, aber die Tatsache, dass kein Mensch dies vorhersagen kann, bedeutet ja noch lange nicht, dass es auch prinzipiell unmöglich ist. Wären sämtliche Einflussfaktoren wie Abwurfgeschwindigkeit und Rotationswinkel, Materialeigenschaften der Münze selbst und solche des Bodens, Masseverteilungen und sogar die Stärke und Richtung des Winds allesamt exakt bekannt, so würden wir mit einigem Aufwand durchaus in der Lage sein, den Ausgang eines solchen Münzwurfs zu prognostizieren. Die physikalischen Formeln hierzu sind jedenfalls alle bekannt.

Wir werden nun ein anderes Szenario betrachten, nur dass wir diesmal sehr wohl den Ausgang der Situation vorhersagen können: Ein Glas wird zerbrechen, wenn es nur aus einer bestimmten Höhe auf einen bestimmten Untergrund fällt. Wir wissen, sogar noch in den Bruchteilen der Sekunde, während das Glas noch fällt: Gleich wird es Scherben geben. Wie sind wir zu dieser eigentlich sehr erstaunlichen Leistung im Stande? Wir haben das betreffende, in diesem Augenblick fallende Glas noch nie vorher zerbrechen sehen und zumindest für die meisten unter uns wird gelten, dass die physikalischen Formeln, welche Glasbruch beschreiben, ein Buch mit sieben Siegeln darstellen. Natürlich kann im Einzelfall das Glas auch einmal „zufällig“ nicht zerbrechen, aber wahrscheinlich ist dieses nicht. Nebenbei bemerkt, „zufällig“ wäre das Nicht-Zerbrechen genauso wenig, da auch dieses Ergebnis physikalischen Gesetzen folgt. Beispielsweise wird die Energie des Aufpralls in diesem Fall günstiger in den Boden übertragen. Woher wissen wir Menschen also in einigen Fällen, was genau als nächstes passieren wird und in anderen, wie beispielsweise beim Münzwurf, nicht?

Die häufigste Erklärung, die Laien in diesem Fall verwenden, ist die Beschreibung des einen Szenarios als „zufällig“ und des anderen als „nicht zufällig“. Wir werden nicht auf die tatsächlich zwar interessanten aber dennoch eher philosophischen Diskussionen zu diesem Thema eingehen, aber wir stellen hier die folgende These auf:

Die allermeisten Prozesse in unserer wahrnehmbaren Umwelt folgen nicht Zufällen. Der Grund für unser Unvermögen, die Prozesse präzise zu beschreiben und zu extrapolieren liegt vielmehr daran, dass wir nicht in der Lage sind, die notwendigen Einflussfaktoren zu erkennen oder zu messen oder diese in die notwendigen Beziehungen zu setzen.

Beim fallenden Glas haben wir die wichtigsten Eigenschaften wie Material, Fallhöhe und Bodenbeschaffenheit schnell erkannt und können innerhalb kürzester Zeit durch Analogieschlüsse aus ähnlichen Erfahrungen bereits eine Schätzung der Wahrscheinlichkeit für Glasbruch abgeben. Beim Münzwurf hingegen schaffen wir genau dieses nicht. Wir können noch so viele Würfe einer Münze betrachten, wir werden es niemals schaffen, bei beliebiger Wurfweise die notwendigen Faktoren schnell genug zu erkennen und entsprechend zu extrapolieren.

Was haben wir also in Gedanken gemacht, als wir die Prognose für den Glaszustand nach dem Aufprall abgegeben haben? Wir haben die Eigenschaften dieses Ereignisses gemessen. Man könnte auch sagen, dass wir Daten gesammelt haben, die den Fall des Glases beschreiben. Blitzschnell haben wir dann einen Analogieschluss durchgeführt, d.h. wir haben gemäß eines Ähnlichkeitsmaßes einen Vergleich mit früheren fallenden Gläsern, Tassen, Porzellanfigürchen oder ähnlichen Gegenständen durchgeführt. Hierzu sind zwei Dinge notwendig, nämlich dass wir die Daten früherer Ereignisse ebenfalls zur Verfügung haben und wir uns im Klaren darüber sind, wie man eine Ähnlichkeit zwischen den aktuellen und den vergangenen Daten überhaupt definiert. Schließlich sind wir in der Lage, eine Schätzung oder Prognose abgegeben, indem wir beispielsweise die ähnlichsten bereits vergangenen Ereignisse betrachtet haben. Ist bei diesen der fallende Gegenstand zerbrochen oder nicht? Dazu müssen wir zunächst mal solche Ereignisse mit größter Ähnlichkeit finden, was eine Art Optimierung darstellt. Wir verwenden hier den Begriff „Optimierung“, da es eigentlich unerheblich ist, ob wir nun eine Ähnlichkeit maximieren oder die Umsätze eines Unternehmens oder beliebiges anderes – in jedem Fall wird die betreffende Größe, also hier die Ähnlichkeit, optimiert. Der beschriebene Analogieschluss liefert uns dann, dass die Mehrzahl der bereits durch uns betrachteten Gläser zerbrochen ist und genau diese Abschätzung wird dann zu unserer Prognose. Dies hört sich vielleicht kompliziert an, aber im Grunde genommen ist diese Art des Analogieschlusses die Basis für beinahe jeden Lernvorgang des Menschen und wird in atemberaubend schneller Zeit durchgeführt.

Das Interessante hieran ist, dass wir soeben als menschliche Data Mining Verfahren tätig waren, denn genau um Fragen wie die Repräsentation von Ereignissen oder Zuständen und die dadurch entstehenden Daten, der Definition von Ähnlichkeiten von Ereignissen und der Optimierung dieser Ähnlichkeiten geht es bei der Datenanalyse üblicherweise.

Beim Münzwurf ist das beschriebene Vorgehen des Analogieschlusses jedoch nicht möglich: es hapert üblicherweise bereits am ersten Schritt und die Daten für Faktoren wie Materialeigenschaften oder Bodenunebenheiten können nicht erfasst werden. Folglich können wir diese auch nicht für spätere Analogieschlüsse bereithalten. Das macht das Ereignis eines Münzwurfs allerdings noch lange nicht zum Zufall, sondern zeigt lediglich, dass wir Menschen nicht in der Lage sind, diese Einflussfaktoren zu messen und den Prozess zu beschreiben. In wieder anderen Fällen sind wir zwar durchaus in der Lage, die Einflussfaktoren zu messen, jedoch gelingt es uns nicht, diese sinnvoll in Beziehung zu setzen, so dass die Berechnung von Ähnlichkeit oder gar die Beschreibung der Prozesse für uns unmöglich ist.

Es ist nun keineswegs so, dass der Analogieschluss die einzige Möglichkeit wäre, aus bereits bekannten Informationen Vorhersagen für neue Situationen abzuleiten. Wird der Beobachter eines fallenden Glases gefragt, woher er wusste, dass das Glas zerbrechen wird, so wird die Antwort häufig Elemente enthalten wie „Immer wenn ich ein Glas habe fallen sehen aus einer Höhe von mehr als 1,5 Metern ist es zerbrochen“. Hier sind zwei Dinge interessant: Der Bezug auf die vergangenen Erfahrungen mittels des Begriffs „immer“ sowie die Ableitung einer Regel aus diesen Erfahrungen:

Wenn der fallende Gegenstand aus Glas ist und die Fallhöhe mehr als 1,5 Meter beträgt, so wird das Glas zerbrechen.

Die Einführung eines Schwellwerts wie 1,5 Meter stellt dabei einen faszinierenden Aspekt dieser Regelbildung dar. Denn obwohl nicht jedes Glas bei größeren Höhen sofort zerbrechen wird und auch nicht bei kleineren Höhen zwingend dem Bruch entfliehen kann, so verwandelt die Einführung dieses Schwellwerts die Regel in eine Daumenregel, die zwar nicht immer, so aber doch in den meisten Fällen zu einer korrekten Einschätzung der Situation führen wird. Anstelle nun also einen direkten Analogieschluss durchzuführen, könnte man sich nun auch dieser Daumenregel bedienen und wird auf diese Weise schnell zu einer Entscheidung über die wahrscheinlichste Zukunft des fallenden Gegenstandes kommen. Analogieschlüsse und die Erstellung von Regeln stellen damit zwei erste Beispiele dar, wie Menschen – und auch Data Mining Verfahren – in der Lage sind, den Ausgang neuer und unbekannter Situationen zu antizipieren.

Unsere Beschreibung dessen, was bei uns im Kopf und auch bei den meisten Data Mining Verfahren im Rechner passiert, offenbart noch eine weitere interessante Einsicht: Der beschriebene Analogieschluss fordert zu keiner Zeit die Kenntnis ir-

gendeiner physikalischen Formel, warum das Glas nun zerbrechen wird. Das gleiche gilt für die oben beschriebene Daumenregel. Selbst ohne also die vollständige (physikalische) Beschreibung eines Vorgangs zu kennen, sind wir und Data Mining Verfahren gleichermaßen bereits in der Lage, eine Abschätzung von Situationen oder gar Prognosen zu generieren. Dabei war ja nicht nur der kausale Zusammenhang selbst unbeschrieben, sondern selbst die Datenerfassung war nur oberflächlich und grob und hat nur wenige Faktoren wie das Material des fallenden Gegenstandes (Glas) und die Fallhöhe (ca. 2m) relativ ungenau abgebildet.

Kausalketten existieren also, ob wir sie nun kennen oder nicht. Im letzteren Fall neigen wir häufig dazu, sie als zufällig zu bezeichnen. Und gleichermaßen ist es erstaunlich, dass selbst für eine unbekannte Kausalkette noch die Beschreibung des weiteren Verlaufs möglich ist, und dies selbst in Situationen, in denen die bisherigen Fakten nur unvollständig und ungenau beschrieben sind.

Dieser Abschnitt hat Ihnen einen Einblick in die Art der Probleme gegeben, denen wir uns in diesem Buch widmen wollen. Wir werden es mit zahlreichen Einflussfaktoren zu tun bekommen, von denen einige gar nicht oder nur unzureichend gemessen werden können. Gleichzeitig sind es oftmals so viele Faktoren, dass wir drohen, den Überblick zu verlieren. Darüber hinaus müssen wir uns noch um die bereits vergangenen Ereignisse kümmern, die wir zur Modellbildung verwenden wollen und deren Anzahl leicht in die Millionen oder Milliarden gehen können. Zu guter Letzt müssen wir uns noch die Frage stellen, ob die Beschreibung des Prozesses das Ziel ist oder ob ein Analogieschluss zur Prognose bereits ausreicht. Und das Ganze muss zudem noch in einer dynamischen Umgebung unter stets wechselnden Bedingungen geschehen – und das am besten möglichst zeitnah. Unmöglich für einen Menschen? Stimmt. Aber eben nicht unmöglich für Data Mining Verfahren.

1.2 Grundbegriffe

Wir werden nun im Folgenden einige Grundbegriffe einführen, die uns die Behandlung der beschriebenen Probleme erleichtern werden. Diese Begriffe werden Sie auch in der Software RapidMiner immer wieder vorfinden, so dass es sich auch für erfahrene Datenanalysten lohnt, die verwendeten Terme kennen zu lernen.

Zunächst einmal können wir feststellen, was die beiden im letzten Abschnitt be-

trachteten Beispiele, der Münzwurf und das fallende Glas, gemeinsam hatten. In unserer Diskussion darüber, ob wir in der Lage sind, das Ende der jeweiligen Situation zu prognostizieren, haben wir festgestellt, dass es auf die möglichst genaue Kenntnis der Einflussfaktoren wie Materialeigenschaften oder Bodenbeschaffenheit ankommt. Und selbst auf die Frage, ob Ihnen dieses Buch weiterhelfen wird, kann man versuchen eine Antwort zu finden, indem man die Eigenschaften von Ihnen, also dem Leser, erfasst und in Einklang bringt mit den Ergebnissen einer Umfrage unter einem Teil der bisherigen Leser. Solche gemessenen Eigenschaften von Lesern könnten dann beispielsweise der Bildungshintergrund der betreffenden Person sein, die Vorliebe für Statistiken, die Präferenzen bei anderen, womöglich ähnlichen Büchern und weitere Merkmale, die wir darüber hinaus noch innerhalb unserer Umfrage messen könnten. Wüssten wir nun von 100 Lesern solche Eigenschaften und hätten von diesen zudem noch die Angabe, ob Ihnen das Buch gefällt oder nicht, so ist das weitere Vorgehen schon beinahe trivial. Wir würden auch Ihnen die Fragen aus unserer Umfrage stellen und auf diese Weise die gleichen Merkmale messen und in Folge, beispielsweise mittels eines Analogieschlusses wie oben beschrieben, eine zuverlässige Prognose ihres persönlichen Geschmacks generieren. „Kunden die dieses Buch gekauft haben, haben auch…“. Das kennen Sie wahrscheinlich schon.

1.2.1 Attribute und Zielattribute

Ob nun Münzen oder andere fallende Gegenstände oder eben auch Menschen, in allen Szenarien steckt wie bereits erwähnt die Frage nach den Eigenschaften oder Merkmalen der jeweiligen Situation. Im Folgenden werden wir stets von **Attributen** sprechen, wenn wir solche beschreibenden Faktoren eines Szenarios meinen. Dies ist auch der Term, der in der Software RapidMiner stets verwendet wird, wenn solche beschreibenden Merkmale auftreten. Die Zahl der Synonyme für diesen Begriff ist hoch, und je nach eigenem Hintergrund werden Ihnen auch schon andere Begriffe anstelle von „Attribut“ begegnet sein, beispielsweise

- Eigenschaft,
- Merkmal (engl. feature),
- Einflussfaktor (engl. influence factor oder auch nur factor),
- Indikator (engl. indicator),

- Variable (engl. variable) oder
- Signal (engl. signal).

Wir haben gesehen, dass die Beschreibung durch Attribute bei Situationen und auch bei Prozessen möglich ist. Dies ist beispielsweise notwendig bei der Beschreibung von technischen Prozessen und hier ist der Gedanke des fallenden Glases gar nicht so weit entfernt. Wenn es möglich ist, den Ausgang einer solchen Situation vorherzusehen, warum dann nicht auch die Qualität eines produzierten Bauteils? Oder den drohenden Ausfall einer Maschine? In gleicher Weise können auch andere Prozesse oder Situationen beschrieben werden, die keinen technischen Bezug haben. Wie kann ich den Erfolg einer Vertriebs- oder Marketingaktion vorhersehen? Welchen Artikel wird ein Kunde als nächstes kaufen? Wie viele Unfälle muss eine Versicherung wohl noch für einen konkreten Kunden oder eine Kundengruppe decken?

Wir werden ein solches Kundenszenario für die Einführung der übrigen wichtigen Begriffe verwenden. Erstens, weil es Menschen bekanntermaßen leichter fällt, Beispiele über andere Menschen zu verstehen. Und zweitens, weil wohl jedes Unternehmen über Informationen, also Attribute, über ihre Kunden verfügt und die meisten Leser die Beispiele daher sofort nachvollziehen können. Die minimal verfügbaren Attribute, die so gut wie jedes Unternehmen über seine Kunden pflegt, sind beispielsweise geographische Angaben und die Information, welche Produkte oder Dienstleistungen der Kunde bereits erworben hat. Sie wären überrascht, welche Vorhersagen bereits aus einer solch kleinen Menge von Attributen möglich sind.

Betrachten wir ein – zugegebenermaßen etwas konstruiertes – Beispiel. Nehmen wir an, dass Sie in einem Unternehmen arbeiten, dass in Zukunft seinen Kunden besser auf ihre Bedürfnisse zugeschnittene Produkte anbieten möchte. Im Rahmen einer Kundenstudie bei nur 100 Ihrer Kunden haben sich einige Bedürfnisse heraus kristallisiert, die immerhin 62 dieser 100 Kunden teilen. Ihre Forschungs- und Entwicklungsabteilung machte sich sofort ans Werk und hat innerhalb kürzester Zeit ein neues Produkt entwickelt, das diesen neuen Bedürfnissen besser gerecht wird. Die meisten der 62 Kunden mit dem entsprechenden Bedürfnisprofil sind von dem Prototypen jedenfalls begeistert, die meisten der übrigen Teilnehmer der Studie zeigen jedoch erwartungsgemäß nur geringes Interesse. Insgesamt haben aber immerhin 54 der 100 Kunden im Rahmen der Studie angegeben, das neue Produkt nützlich zu finden. Der Prototyp wird also als Erfolg bewertet und geht

in Produktion – nur stellt sich nun die Frage, wie Sie aus ihren Bestandskunden oder auch aus anderen potentiellen Kunden genau diejenigen heraussuchen, bei denen die dann folgenden Marketing- und Vertriebsbemühungen auch den größten Erfolg versprechen. Sie möchten also ihre Effizienz in diesem Bereich optimieren und dazu gehört insbesondere, solche Bemühungen von vorneherein auszuschließen, die ohnehin nur mit geringer Wahrscheinlichkeit zu einem Kauf führen. Aber wie macht man das? Das Bedürfnis nach alternativen Lösungen und damit das Interesse an dem neuen Produkt hat sich ja im Rahmen der Kundenstudie auf einer Teilmenge Ihrer Kunden ergeben. Der Aufwand, diese Studie flächendeckend durchzuführen ist viel zu hoch und verbietet sich daher von selbst. Und genau hier kann Data Mining helfen. Betrachten wir zunächst eine mögliche Auswahl von Attributen über ihre Kunden:

- Name
- Adresse
- Branche
- Subbranche
- Zahl der Mitarbeiter
- Anzahl der Käufe in Produktgruppe 1
- Anzahl der Käufe in Produktgruppe 2
- ...

Die Anzahl der Käufe in den unterschiedlichen Produktgruppen meint hier die Transaktionen in Ihren Produktgruppen, die Sie in der Vergangenheit mit diesem Kunden bereits getätigt haben. Natürlich können in Ihrem Fall auch mehr oder weniger oder auch ganz andere Attribute vorhanden sein, aber das soll an dieser Stelle keine Rolle spielen. Nehmen wir an, dass Ihnen die Informationen über diese Attribute für jeden Ihrer Kunden zur Verfügung ständen. Dann gibt es aber noch ein Attribut, welches wir für unser ganz konkretes Szenario betrachten können: Die Tatsache nämlich, ob dem Kunden der Prototyp gefällt oder eben auch nicht. Dieses Attribut steht natürlich nur für die 100 Kunden aus der Studie zur Verfügung, für die anderen ist die Information über dieses Attribut schlicht unbekannt. Trotzdem nehmen wir das Attribut ebenfalls mit in die Liste unserer Attribute auf:

- *Prototyp positiv aufgenommen?*
- Name
- Adresse
- Branche
- Subbranche
- Zahl der Mitarbeiter
- Anzahl der Käufe in Produktgruppe 1
- Anzahl der Käufe in Produktgruppe 2
- ...

Nehmen wir an, sie haben insgesamt tausende von Kunden, so können Sie lediglich bei 100 von diesen eine Angabe darüber machen, ob der Prototyp positiv bewertet wurde oder nicht. Bei den anderen, wissen Sie dies noch nicht – aber Sie würden es gerne wissen! Das Attribut „Prototyp positiv aufgenommen?“ nimmt also eine Sonderrolle ein, da es jeden Ihrer Kunden in Bezug zu der augenblicklichen Fragestellung kennzeichnet. Wir nennen dieses besondere Attribut daher auch **Label**, da es wie ein Markenlabel an einem Hemd oder auch ein Notizzettel an einer Pinnwand an ihren Kunden haftet und diese kennzeichnet. Unter den Namen „Label“ werden Sie Attribute, die diese spezielle Rolle annehmen, auch in RapidMiner wiederfinden. Das Ziel unserer Bemühungen ist ja, für die Gesamtmenge aller Kunden dieses konkrete Attribut auszufüllen. Daher werden wir in diesem Buch auch oft von **Zielattribut** anstelle des Begriffs „Label“ sprechen. In der Literatur werden Sie auch häufig den Begriff Zielvariable entdecken, der ebenfalls das gleiche meint.

1.2.2 Konzepte und Beispiele

Die oben eingeführte Strukturierung von Eigenschaften ihrer Kunden durch Attribute hilft uns schon einmal, das gestellte Problem etwas analytischer angehen zu können. Wir haben auf diese Weise nämlich sicher gestellt, dass jeder Ihrer Kunden auf die gleiche Art und Weise repräsentiert wird. Wir haben im gewissen Sinne den Typ oder das **Konzept** „Kunde“ definiert, welches sich deut-

lich von anderen Konzepten wie beispielsweise „fallende Gegenstände“ dadurch unterscheidet, dass Kunden typischerweise keine Materialeigenschaften besitzen und fallende Gegenstände nur selten in Produktgruppe 1 einkaufen werden. Es ist wichtig, dass Sie für jedes der Probleme in diesem Buch – oder auch solchen in Ihrer eigenen Praxis – zunächst definieren, mit welchen Konzepten Sie es eigentlich zu tun haben und durch welche Attribute diese definiert werden.

Oben haben wir implizit durch die Angabe der Attribute Name, Adresse, Branche usw. und insbesondere der Angabe der Kauftransaktionen in den einzelnen Produktgruppen definiert, dass durch diese Attribute Objekte des Konzepts „Kunde“ beschrieben werden. Nun ist dieses Konzept bisher relativ abstrakt geblieben und noch nicht mit Leben gefüllt. Wir wissen zwar nun, auf welche Weise wir Kunden beschreiben können, haben dies allerdings bisher noch nicht für konkrete Kunden durchgeführt. Betrachten wir beispielsweise die Attribute des folgenden Kunden:

- *Prototyp positiv aufgenommen: ja*
- Name: *Müller Systemtechnik GmbH*
- Adresse: *Meisenstr. 7, Böblingen*
- Branche: *Industrie*
- Subbranche: *Rohrbiegemaschinen*
- Zahl der Mitarbeiter: > 1000
- Anzahl der Käufe in Produktgruppe 1: *5*
- Anzahl der Käufe in Produktgruppe 2: *0*
- ...

Wir sagen, dass dieser konkrete Kunde ein **Beispiel** für unser Konzept „Kunde“ ist. Jedes Beispiel kann durch seine Attribute charakterisiert werden und besitzt für diese Attribute konkrete Werte, die mit denen anderer Beispiele verglichen werden können. In dem oben beschriebenen Fall, handelt es sich mit der Müller Systemtechnik GmbH darüber hinaus noch um das Beispiel eines Kunden, welches an unserer Studie teilgenommen hat. Daher liegt für unser Zielattribut „Prototyp positiv aufgenommen?“ ein Wert vor. Die Müller Systemtechnik war zufrieden und hat hier ein „ja“ als Attributwert, daher sprechen wir auch von einem **positiven Beispiel**. Folgerichtig gibt es auch negative Beispiele und solche

Beispiele, bei denen wir gar keine Aussage über das Zielattribut machen können.

1.2.3 Attributrollen

Wir haben nun schon zwei verschiedene Arten von Attributen kennen gelernt, nämlich solche, die die Beispiele einfach nur beschreiben und solche, die die Beispiele gesondert kennzeichnen. Attribute können also verschiedene Rollen annehmen. Wir haben bereits die Rolle „Label“ eingeführt für Attribute, welche die Beispiele in irgendeiner Weise kennzeichnen und welche es für neue Beispiele, die noch nicht derart gekennzeichnet sind, vorherzusagen gilt. In unserem oben beschriebenen Szenario beschreibt das Label – sofern vorhanden – nach wie vor die Eigenschaft, ob der Prototyp positiv aufgenommen wurde.

Gleichermaßen gibt es beispielsweise Rollen, bei denen das zugehörige Attribut zur eindeutigen Identifikation des betreffenden Beispiels dient. In diesem Fall nimmt das Attribut die Rolle einer Identifizierung ein und wird kurz **ID** genannt. Mit dieser Rolle finden Sie solche Attribute auch in der Software RapidMiner gekennzeichnet. In unserem Kundenszenario könnte das Attribut „Name“ die Rolle einer solchen Identifikation einnehmen.

Es gibt noch weitere Rollen, wie beispielsweise solche, bei denen das Attribut das Gewicht des Beispiels hinsichtlich des Labels bezeichnet. In diesem Fall trägt die Rolle den Namen „Gewicht“ oder **Weight**. Attribute ohne besondere Rolle, also solche, die die Beispiele einfach nur beschreiben, nennen wir auch **reguläre** Attribute und lassen die Rollenbezeichnung in den meisten Fällen einfach weg. Im Übrigen steht es Ihnen in RapidMiner frei, auch eigene Rollen zu vergeben und somit Ihre Attribute gesondert in Ihrer Bedeutung zu kennzeichnen.

1.2.4 Wertetypen

Neben den verschiedenen Rollen eines Attributs gibt es noch eine zweite Eigenschaft von Attributen, die eine genauere Betrachtung verdient. Das Beispiel der Müller Systemtechnik oben hat für die verschiedenen Attribute die jeweiligen Werte definiert, beispielsweise „Müller Systemtechnik GmbH“ für das Attribut „Name“ und den Wert „5“ für die Anzahl der bisherigen Käufe in Produktgruppe 1. Für das Attribut „Name“ handelt es sich bei dem konkreten Wert für dieses Beispiel also gewissermaßen um nahezu beliebigen Freitext, beim Attribut „Anzahl

der Käufe in Produktgruppe 1“ wiederum muss die Angabe einer Zahl entsprechen. Die Angabe, ob die Werte eines Attribut nun als Text oder Zahl vorliegen müssen, nennen wir den **Wertetyp** (engl.: **Value Type**) eines Attributs.

In späteren Kapiteln werden wir viele verschiedene Wertetypen kennenlernen und sehen, wie sich diese auch in andere Typen transformieren lassen. Für den Augenblick reicht uns die Erkenntnis, dass es verschiedene Wertetypen für Attribute gibt und dass wir im Fall von Freitext von dem Wertetyp **Text**, im Fall von Zahlen von dem Wertetyp **Numerisch** oder englisch **Numerical** und im Fall von nur wenigen möglichen Werten – wie etwas bei den beiden Möglichkeiten „ja“ und „nein“ beim Zielattribut – von dem Wertetyp **Nominal** sprechen. Bitte beachten Sie, dass im obigen Beispiel die Zahl der Mitarbeiter, obwohl ja eigentlich vom numerischen Typ, eher als nominal definiert werden würde, da statt einer genauen Angabe wie 1250 Mitarbeiter eine Größenklasse, nämlich „> 1000“, verwendet wurde.

Die folgende Tabelle gibt Ihnen einen Überblick über alle von RapidMiner unterstützten Wertetypen:

Wertetyp	**RapidMiner-Name**	**Verwendung**
Nominal	nominal	Kategorielle nicht-numerische Werte, meist für endliche Mengen von unterschiedlichen Ausprägungen verwendet
Numerische Werte	numeric	Allgemein für numerische Werte
Ganzzahlen	integer	Ganze Zahlen, positiv und negativ
Reelle Zahlen	real	Reelle Zahlen, positiv und negativ
Text	text	Beliebiger Freitext ohne Struktur
2-wertig Nominal	binominal	Spezialfall von Nominal, bei dem nur zwei unterschiedliche Werte erlaubt sind
mehrwertig Nominal	polynominal	Spezialfall von Nominal, bei dem mehr als zwei unterschiedliche Werte erlaubt sind
Date Time	data_time	Datum zusammen mit Zeit
Datum	date	Nur Datum
Zeit	time	Nur Zeit

1.2.5 Daten und Metadaten

Wir wollen unsere Ausgangssituation noch einmal zusammenfassen. Wir haben ein **Konzept** „Kunde“ vorliegen, welches wir mit einer Reihe von **Attributen** beschrieben wird:

- *Prototyp positiv aufgenommen? Label; Nominal*
- Name: *Text*
- Adresse: *Text*
- Branche: *Nominal*
- Subbranche: *Nominal*
- Zahl der Mitarbeiter: *Nominal*
- Anzahl der Käufe in Produktgruppe 1: *Numerisch*
- Anzahl der Käufe in Produktgruppe 2: *Numerisch*
- ...

Das Attribut „Prototyp positiv aufgenommen?“ hat eine besondere **Rolle** unter den Attributen, es handelt sich hierbei um unser **Zielattribut**. Das Zielattribut hat den **Wertetyp Nominal**, was bedeutet, dass nur relativ wenige Ausprägungen (in diesem Fall „ja“ und „nein“) angenommen werden können. Genau genommen ist es sogar binominal, da nur zwei verschiedene Ausprägungen erlaubt sind. Die übrigen Attribute haben alle keine gesonderte Rolle, d.h. sie sind **regulär**, und haben entweder den Wertetyp **Numerisch** oder **Text**. Die folgende Definition ist sehr wichtig, da sie für eine erfolgreiche professionelle Datenanalyse eine zentrale Rolle spielt:

Diese Menge an Informationen, die ein Konzept beschreiben, nennen wir auch Metadaten, da sie Daten über die eigentlichen Daten darstellen.

Für unser Konzept „Kunde“ hat unser fiktives Unternehmen jede Menge **Beispiele**, nämlich die Informationen, die das Unternehmen zu den einzelnen Attributen in seiner Kundendatenbank gespeichert hat. Das Ziel ist nun, aus den Beispielen, für die eine Information über das Zielattribut vorliegt, eine Prognoseanweisung zu generieren, welche uns für die übrigen Kunden vorhersagt, ob diese mit höherer

Wahrscheinlichkeit den Prototypen eher positiv aufnehmen würden oder ablehnen. Die Suche nach solch einer Prognoseanweisung ist eine der Aufgaben, die man mit Hilfe von Data Mining lösen kann.

Hierzu ist es jedoch wichtig, dass die Informationen zu den Attributen der einzelnen Beispiele in einer geordneten Form vorliegen, damit Data Mining Verfahren auf diese mittels eines Rechners zugreifen können. Was läge hier näher als eine Tabelle? Jedes der Attribute definiert eine Tabellenspalte und jedes Beispiel mit den verschiedenen Attributwerten entspricht einer Zeile dieser Tabelle. Für unser Szenario könnte dies beispielsweise wie in Tabelle 1.1 aussehen.

Eine solche Tabelle nennen wir **Beispielmenge** oder englisch **Example Set**, da diese Tabelle die Daten für alle Attribute unserer Beispiele enthält. Im Folgenden und auch innerhalb von RapidMiner werden wir die Begriffe **Daten, Datensatz** und **Beispielmenge** synonym verwenden. Stets ist in diesem Fall eine Tabelle mit den entsprechenden Einträgen für die Attributwerte der aktuellen Beispiele gemeint. Solche Datentabellen sind es auch, die der Datenanalyse oder dem Data Mining ihren Namen geliehen haben. Merke:

Daten *beschreiben die Objekte eines Konzepts,* ***Metadaten*** *beschreiben die Eigenschaften eines Konzepts (und damit auch der Daten).*

Die meisten Data Mining Verfahren erwarten, dass die Beispiele genau in solch einer Attributwertetabelle gegeben werden. Dies ist hier glücklicherweise der Fall und wir können uns weitere Transformationen der Daten sparen. In der Praxis sieht dies jedoch ganz anders aus und der Großteil des Arbeitsaufwandes bei einer Datenanalyse wird für die Übertragung der Daten in ein für das Data Mining geeignetes Format aufgewendet. Diese Transformationen werden daher intensiv in späteren Kapiteln behandelt.

1.2.6 Modellierung

Nachdem wir die Daten über unsere Kunden in einem gut strukturierten Format vorliegen haben, können wir nun also endlich mittels eines Data Mining Verfahrens die unbekannten Werte unseres Zielattributs durch die Prognose des wahrscheinlichsten Werts ersetzen. Hierbei stehen uns zahlreiche Verfahren zur Verfügung, von denen viele, wie auch der eingangs beschriebene Analogieschluss oder das Generieren von Daumenregeln, dem menschlichen Verhalten nachemp-

Prototyp positiv aufgenommen?	*Name*	*Adresse*	*Branche*	*Unterbranche*	*Zahl Mitarbeiter*	*Anzahl Käufe Gruppe 1*	*Anzahl Käufe Gruppe 2*	...
ja	Müller Systemtechnik GmbH	Meisenstr. 7, Böblingen	Industrie	Rohrbiegemaschinen	> 1000	5	0	...
?	Meier Papier	Taubenweg 6, Coburg	IT	Telekommunikation	600–1000	3	7	...
nein	Schulze & Nagel	Amselallee 5, Homberg	Handel	Textilien	<100	1	11	...
...	...	...	...	...	...	...	...	...

Tabelle 1.1: Ein Beispielszenario

funden sind. Die Anwendung eines Data Mining Verfahrens nennen wir **modellieren** und das Ergebnis eines solchen Verfahrens, also die Prognoseanweisung, ist ein **Modell**. Genau wie Data Mining insgesamt für unterschiedliche Fragestellungen angewendet werden kann, so gilt dieses auch für Modelle. Sie können leicht verständlich sein und Ihnen auf einfache Weise die zu Grunde liegenden Prozesse erklären. Oder sie können sich gut zur Prognose für unbekannte Situationen einsetzen lassen. Manchmal gilt auch beides, wie beispielsweise bei dem folgenden Modell, welches ein Data Mining Verfahren für unser Szenario geliefert haben könnte:

„Wenn der Kunde aus städtischen Gebieten kommt, mehr als 500 Mitarbeiter hat und mindestens 3 Käufe in Produktgruppe 1 getätigt wurden, so ist die Wahrscheinlichkeit hoch, dass dieser Kunde sich für das neue Produkt interessiert."

Ein solches Modell ist leicht verständlich und gibt unter Umständen tiefere Einsichten in die zu Grunde liegenden Daten und Entscheidungsprozesse Ihrer Kunden. Und es handelt sich darüber hinaus um ein operationales Modell, also um ein Modell welches direkt zur Prognose für weitere Kunden eingesetzt werden kann. Die Firma „Meier Papier" beispielsweise erfüllt die Bedingungen der obigen Regel und wird sicher daher ebenfalls für das neue Produkt interessieren – jedenfalls gilt dieses mit höherer Wahrscheinlichkeit. Ihr Ziel wäre also erreicht und Sie hätten mit Hilfe von Data Mining ein Modell generiert, welches Sie zur Steigerung Ihrer Marketingeffizienz einsetzen könnten: Statt nun einfach alle Bestandskunden und sonstige Kandidaten blind zu kontaktieren, können Sie ihre Vertriebsbemühungen nun auf die vielversprechenden Kunden konzentrieren und haben so bei geringerem Aufwand eine wesentlich höhere Erfolgsquote. Oder Sie gehen sogar noch einen Schritt weiter und analysieren, für welche Kunden wohl welche Vertriebskanäle die besten Ergebnisse liefern werden.

In den folgenden Kapiteln werden wir uns weiteren Anwendungen für Data Mining widmen und ganz nebenbei die Überführung von Konzepten wie Kunden, Geschäftsprozessen oder Produkten in Attribute, Beispiele und Datensätze trainieren. Dies schult den Blick für weitere Anwendungsmöglichkeiten ungemein und wird Ihnen später das Analystenleben deutlich erleichtern. Zunächst wollen wir uns jedoch kurz mit RapidMiner beschäftigen und eine kleine Einführung in die Bedienung geben, damit Sie die nachfolgenden Beispiele auch direkt umsetzen können.

2 Design von Analyseprozessen mit RapidMiner

Die Analyse großer Datenmengen mit Methoden des Data Mining wird gemeinhin als ein Feld für Spezialisten betrachtet. Diese erstellen mit häufig sündhaft teuren Softwarelösungen mehr oder weniger komplexe Analyseprozesse, um beispielsweise drohende Kündigungen oder die Verkaufszahlen eines Produkts zu prognostizieren. Der wirtschaftliche Nutzen liegt auf der Hand, und so galt lange Zeit, dass die Anwendung von Data Mining Softwareprodukten auch mit hohen Kosten für Softwarelizenzen und den auf Grund der Komplexität der Materie oft notwendigen Support verbunden war. Dass Softwarelösungen für Data Mining jedoch nicht zwingend teuer oder schwer zu bedienen sein müssen, daran dürfte spätestens seit der Entwicklung der Open Source Software RapidMiner wohl niemand mehr ernsthaft zweifeln.

Begonnen wurde die Entwicklung von RapidMiner unter dem Namen „Yet Another Learning Environment“ (YALE) am Lehrstuhl für künstliche Intelligenz der Universität Dortmund unter der Leitung von Prof. Dr. Katharina Morik. Mit der Zeit wurde die Software immer ausgereifter, mehr als eine halbe Million Downloads wurden seit dem Entwicklungsstart im Jahre 2001 verzeichnet. Unter den vielen Tausend Anwendern waren auch viele Unternehmen, welche nach einem Partner mit entsprechender Data Mining Kompetenz für Dienstleistungen und Projekte suchten. Diesem Bedarf folgend, wurde von den RapidMiner-Entwicklern das Unternehmen Rapid-I gegründet, welches heute auch für die Weiterentwicklung und Wartung der Software verantwortlich ist. Im Zuge der Unternehmensgründung wurde die Software YALE ihrer neuen Bedeutung ent-

sprechend in RapidMiner umbenannt. Damit befinden sich RapidMiner und das dahinter stehende Unternehmen Rapid-I auf einem guten Wege: Rapid-I erreichte den vierten Platz beim nationalen Start-Up Wettbewerb „start2grow“ und gewann bei Europas höchstdotiertem IT-Wettbewerb „Open Source Business Award“ den ersten Preis. RapidMiner selbst wurde auf dem bekannten Data Mining Portal „KDnuggets“ bereits zum dritten Mal in Folge zur meistverwendeten Open Source Data Mining Lösung gewählt – und auch insgesamt machte RapidMiner mit einem knappen zweiten Platz unter den mehr als 30 auch proprietären Lösungen eine mehr als gute Figur.

2.0.1 Flexibilität und Funktionsvielfalt

Was genau macht RapidMiner aber zur weltweit führenden Open Source Data Mining Software? Gemäß einer unabhängigen Vergleichsstudie der TU Chemnitz, die beim internationalen Data Mining Cup 2007 (DMC-2007) vorgestellt wurde, schneidet RapidMiner unter den wichtigsten Open Source Data Mining Tools sowohl hinsichtlich der Technologie als auch der Anwendbarkeit am besten ab. Dies spiegelt auch den Fokus der Entwicklungsarbeit wieder, der stets auf eine benutzerfreundliche Kombinierbarkeit der aktuellsten sowie der bewährten Data Mining Techniken abzielte.

Diese Kombinationsfreudigkeit verschafft RapidMiner eine hohe Flexibilität bei der Definition von Analyseprozessen. Wie wir im Folgenden sehen werden, können *Prozesse* aus einer großen Zahl von nahezu beliebig schachtelbaren *Operatoren* erzeugt und schließlich durch sogenannte Operator Trees beziehungsweise durch einen Prozessgraphen (*Flow Design*) repräsentiert werden. Der Prozessaufbau wird intern durch XML beschrieben und mittels einer graphischen Benutzeroberfläche entwickelt. Im Hintergrund prüft RapidMiner ständig den gerade entwickelten Prozess auf Syntaxkonformität und gibt automatisch Vorschläge für den Problemfall. Dies wird ermöglicht durch eine die sogenannte *Metadaten-Transformation*, welche bereits zur Design-Zeit die zu Grunde liegenden Metadaten so transformiert, dass die Form des Ergebnisses bereits absehbar ist und bei unpassenden Operatorkombinationen Lösungen aufgezeigt werden können (*Quick Fixes*). Weiterhin bietet RapidMiner dem Analysten die Möglichkeit, *Breakpoints* zu definieren und damit praktisch jedes Zwischenergebnis inspizieren zu können. Gelungene Kombinationen von Operatoren können zusammen gefasst werden in *Building Blocks* und stehen damit in späteren Prozessen erneut zur Verfügung.

Damit kombinieren die Prozesse von RapidMiner die Mächtigkeit von Entwicklungsumgebungen, wie man sie von Programmiersprachen kennt, mit der Einfachheit von visueller Programmierung. Das modulare Vorgehen hat zudem den Vorteil, dass auch die internen Analyseabläufe genauestens geprüft und ausgenutzt werden können. Analysten können so beispielsweise auch in die einzelnen Teilschritte einer Kreuzvalidierung hineinsehen oder den Effekt der Vorverarbeitung ebenfalls evaluieren – was mit anderen Lösungen typischerweise nicht möglich ist und oftmals in zu optimistischen Fehlerabschätzungen resultiert.

Insgesamt beinhaltet RapidMiner mehr als 500 Operatoren für alle Aufgaben der professionellen Datenanalyse, d.h. Operatoren für *Ein-* und *Ausgabe* sowie der *Datenverarbeitung (ETL)*, *Modellierung* und anderen Aspekten des Data Mining. Aber auch Methoden des *Text Mining*, *Web Mining*, der automatischen Stimmungsanalyse aus Internet-Diskussionsforen (*Sentiment Analysis*, Opinion Mining) sowie der *Zeitreihenanalyse* und *-prognose* stehen dem Analysten zur Verfügung. Zusätzlich beinhaltet RapidMiner mehr als 20 Verfahren, auch hochdimensionale Daten und Modelle zu *visualisieren.* Darüber hinaus wurden auch alle Lernverfahren und Gewichtungsfaktoren der Weka Toolbox vollständig und nahtlos in RapidMiner integriert, so dass zu dem bereits enormen Funktionsumfang von RapidMiner auch noch einmal der vollständige Funktionsumfang des gerade in der Forschung ebenfalls weit verbreiteten Weka kommt.

2.0.2 Skalierbarkeit

Im Oktober 2009 erschien die Version 4.6 von RapidMiner und Ende 2009 dann endlich die vollständig neu gestaltete Version 5.0. Die Stoßrichtung wird in diesen beiden Versionen mehr als deutlich: zusätzlich zur großen Funktionsvielfalt liegt der Hauptfokus auf eine Optimierung hinsichtlich der Skalierbarkeit auch auf große Datenmengen. Schon immer war eine der Haupteigenschaften von RapidMiner ein Konzept ähnlich zu dem von relationalen Datenbanken, welches verschiedene *Sichten* auf Datenquellen ermöglicht. Dieses Konzept hat RapidMiner weiter verfeinert und bietet nun die Möglichkeit, eine Vielzahl solcher Sichten so zu kombinieren, dass die Daten *on-the-fly* transformiert und Datenkopien weitestgehend unnötig werden. Hierdurch erreicht RapidMiner einen im Vergleich oftmals deutlich niedrigeren Speicherverbrauch und kann – eine entsprechende Konfiguration von RapidMiner und der Analyseprozesse vorausgesetzt – auch mit mehreren 100 Millionen Datensätzen spielend leicht umgehen.

Weitere Neuerungen wie die verbesserten Lift Charts von RapidMiner unterstützen die Optimierung von Direct-Mailing- und Marketing-Kampagnen, die Kündigerprävention (Churn Reduction), die Erhöhung der Kundenbindung und die Kosten-Nutzen-optimierte Neukundengewinnung. Erweiterte Pivotisierungen, neue Aggregationsfunktionen, eine umfangreiche Datums- und Zeitbehandlung, die vereinfachte funktionsbasierte Konstruktion neuer Attribute, optimierte Wizards unter anderem für die automatische Optimierung von Data Mining Prozessparametern sowie neue Visualisierungen mit Zooming und Panning ermöglichen ebenfalls verbesserte Analysen und Datentransformationen und erleichtern die Bedienung zudem enorm. Die wesentlichsten Neuerungen der neuen Version 5 von RapidMiner ist jedoch die vollständige Überarbeitung der graphischen Benutzeroberfläche, die statt lediglich des Operatorbaums nun auch die expliziten Datenflüsse anzeigt und zudem auf Basis des nun integrierten Repositories auch die Metadaten-Transformation während der Design-Zeit unterstützt.

2.0.3 Eine Frage des Formats

Ein weiterer Schwerpunkt von RapidMiner ist die hohe Konnektivität zu den verschiedensten Datenquellen wie z.B. Oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL und Ingres, dem Zugriff auf Excel-, Access- und SPSS-Dateien sowie zahlreichen anderen Datenformaten. Zusammen mit den hunderten Operatoren zur Datenvorverarbeitung lässt sich RapidMiner neben der Datenanalyse damit auch hervorragend zur Datenintegration und -transformation (ETL) einsetzen.

Und auch bei der Software selbst hat der Anwender die Wahl aus verschiedenen Formaten. RapidMiner gibt es einmal in der freien RapidMiner Community Edition, welche jederzeit und kostenlos von der Website heruntergeladen werden kann und in der Enterprise Edition, welche die Vorteile der freien Community Edition mit einem vollständigen professionellen Support mit garantierten Antwortzeiten kombiniert.

2.1 Installation und Erstes Repository

Bevor wir mit RapidMiner arbeiten können, müssen Sie die Software natürlich erst einmal herunterladen und installieren. Sie finden sie auf der Webseite des

Herstellers Rapid-I im Downloadbereich unter

http://www.rapid-i.com

Laden Sie das passende Installationspaket für Ihr Betriebssystem herunter und installieren Sie RapidMiner gemäß den Anweisungen auf der Webseite. Es werden alle gängigen Windowsversionen genauso unterstützt wie Macintosh, Linux oder Unix Systeme. Beachten Sie bitte, dass auf den letztgenannten eine aktuelle Java Runtime mit mindestens Version 6 erforderlich ist.

Sollten Sie RapidMiner zum ersten Mal starten, werden Sie dazu aufgefordert, ein neues Repository anzulegen. Wir werden uns zunächst auf ein lokales Repository auf Ihrem Rechner beschränken, später können Sie dann auch Repositories im Netzwerk definieren, die Sie darüber hinaus auch mit anderen Analysten teilen können:

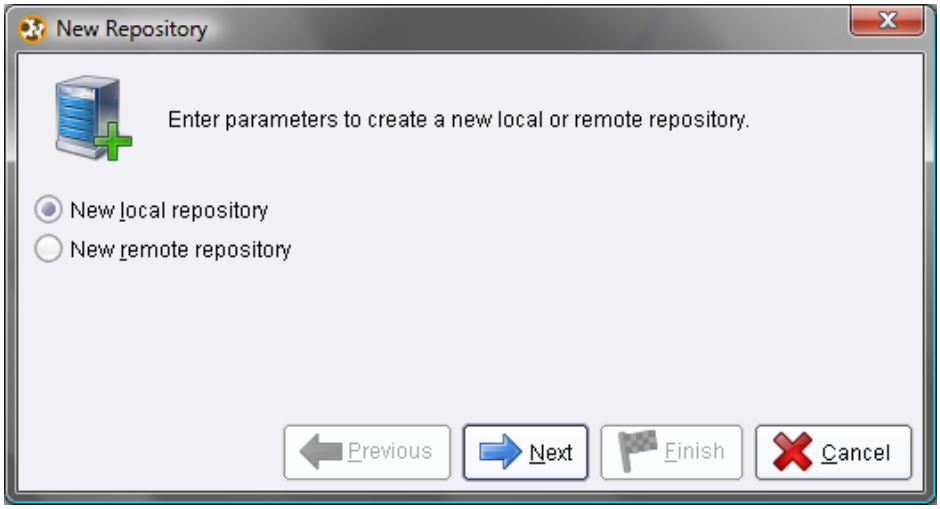


Abbildung 2.1: Legen Sie beim ersten Programmstart zunächst ein lokales Repository auf Ihrem Rechner an.

Für ein lokales Repository brauchen Sie lediglich einen Namen festzulegen (Alias) und ein beliebiges Verzeichnis auf Ihrer Festplatte zu definieren. Sie können das Verzeichnis direkt mittels eines Klicks auf das Ordner-Icon rechts wählen. Es empfiehlt sich, im dann aufkommenden Dateidialog ein neues Verzeichnis an einem Ihnen angenehmen Ort anzulegen und dieses neue Verzeichnis dann als Basis für Ihr lokales Repository zu verwenden. Dieses Repository dient als zentraler

Speicherort für Ihre Daten und Analyseprozesse und wird Sie in der nächsten Zeit begleiten.

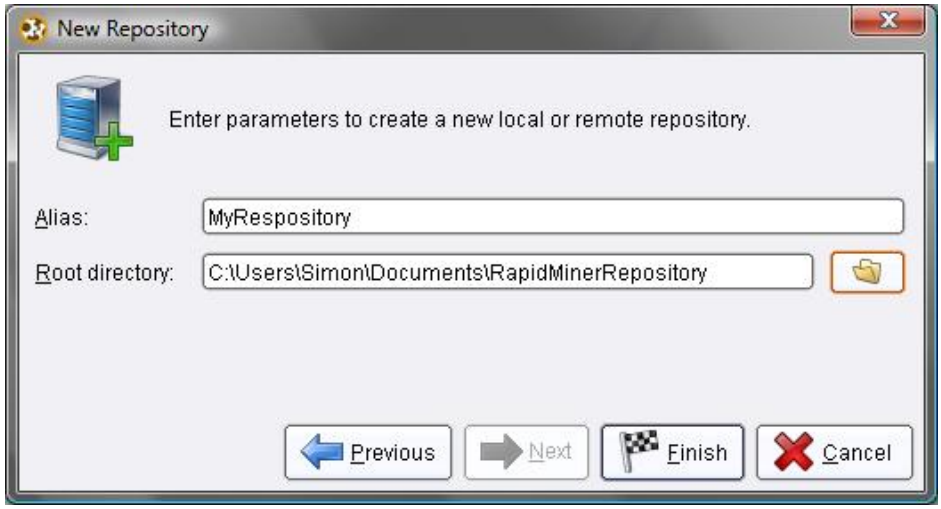


Abbildung 2.2: Definition eines neuen lokalen Repositories zu Speicherung Ihrer Daten und Analyseprozesse. Es empfiehlt sich, ein neues Verzeichnis als Basis anzulegen.

2.2 Perspektiven und Views

Nach der Wahl des Repositories werden Sie in der sogenannten *Welcome-Perspektive* begrüßt (Abbildung 2.3).

Der untere Bereich zeigt aktuelle *Nachrichten* rund um RapidMiner an, sofern Sie eine Verbindung zum Internet besitzen. Die Liste in der Mitte zeigt die *zuletzt bearbeiteten Analyseprozesse*. Dies ist praktisch, wenn Sie einen dieser Prozesse weiter bearbeiten oder ausführen möchten. Sie können einen Prozess aus dieser Liste einfach mittels eines Doppelklicks zur Bearbeitung oder Durchführung öffnen. Der oberste Bereich schließlich zeigt typische Aktionen, die Sie als Analyst nach dem Start von RapidMiner häufig durchführen. Im Einzelnen sind dies

1. *New:* Startet einen neuen Analyseprozess. Zunächst müssen Sie einen Ort und einen Namen innerhalb des Prozess- und Datenrepositories definieren und können dann mit dem Design eines neuen Prozesses starten.

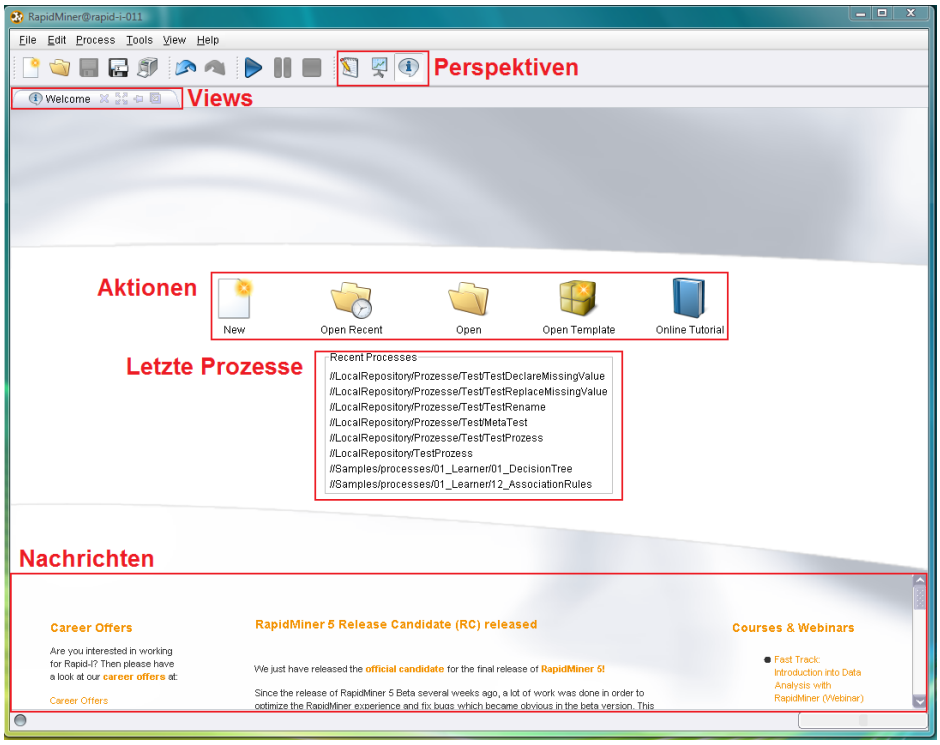

File Edit Process Tools View Help
Perspektiven
Welcome
Views
Aktionen
New
Open Recent
Open
Open Template
Online Tutorial
Letzte Prozesse
Recent Processes
//LocalRepository/Prozesse/Test/TestDeclareMissingValue
//LocalRepository/Prozesse/Test/TestReplaceMissingValue
//LocalRepository/Prozesse/Test/TestRename
//LocalRepository/Prozesse/Test/MetaTest
//LocalRepository/Prozesse/Test/TestProzess
//LocalRepository/TestProzess
//Samples/processes/01_Learner/01_DecisionTree
//Samples/processes/01_Learner/12_AssociationRules
Nachrichten
Career Offers
Are you interested in working for Rapid-I? Then please have a look at our career offers at:
Career Offers
RapidMiner 5 Release Candidate (RC) released
We just have released the official candidate for the final release of RapidMiner 5!
Since the release of RapidMiner 5 Beta several weeks ago, a lot of work was done in order to optimize the RapidMiner experience and fix bugs which became obvious in the beta version. This
Courses & Webinars
Fast Track: Introduction into Data Analysis with RapidMiner (Webinar)

Abbildung 2.3: Welcome-Perspektive von RapidMiner.

2. *Open Recent:* Öffnet den Prozess, der in der Liste unterhalb der Aktionen ausgewählt ist. Sie können alternativ diesen Prozess auch mittels eines Doppelklicks innerhalb der Liste öffnen. In jedem Fall wechselt RapidMiner auch hier danach automatisch in die Design-Perspektive.

3. *Open:* Öffnet den Repository-Browser und erlaubt die Auswahl eines Prozesses zum Öffnen innerhalb der Prozess Design-Perspektive.

4. *Open Template:* Zeigt eine Auswahl von verschiedenen vordefinierten Analyseprozessen, die innerhalb weniger Klicks konfiguriert werden können.

5. *Online Tutorial:* Startet ein Tutotial, welches direkt innerhalb von RapidMiner verwendet werden kann und anhand einer Auswahl von Analyseprozessen in einige Konzepte des Data Mining einführt. Empfehlenswert, wenn Sie bereits Grundwissen im Bereich Data Mining haben und mit der grundlegenden Bedienung von RapidMiner vertraut sind.

In der Toolbar im obersten Bereich von RapidMiner finden Sie am rechten Rand drei Icons, welche zwischen den einzelnen Perspektiven von RapidMiner umschalten. Eine *Perspektive* besteht aus einer frei konfigurierbaren Auswahl von einzelnen Elementen der Oberfläche, den sogenannten *Views*. Diese können zudem noch beliebig angeordnet werden.

In der Welcome-Perspektive gibt es zumindest voreingestellt nur einen einzigen View, nämlich den Willkommensschirm, den Sie gerade vor sich sehen. Sie können weitere Views aktivieren, indem Sie das Menü „View“ aufrufen:

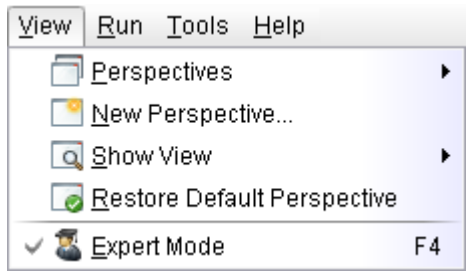


Abbildung 2.4: View Menü.

Im Unterpunkt „Show View“ finden Sie alle verfügbaren Views von RapidMiner. Views, die in der aktuellen Perspektive gerade sichtbar sind, werden durch ein Häkchen gekennzeichnet. Schalten Sie durch eine Auswahl einen weiteren View ein, beispielsweise den View mit dem Namen „Log“. Sie sehen in Abbildung 2.5 nun, dass in der Wilkommensperspektive ein zweiter View mit diesem Namen hinzugefügt wurde.

Sie sehen oben nun den bereits bekannten Welcome View und unten den neuen Log View. Wenn Sie die Maus nun in den markierten Bereich dazwischen bewegen, so ändert der Mauszeiger seine Form und zeigt an, dass Sie durch Draggen, also durch Ziehen des Zeigers bei gedrückter Taste, die Größen der Views ändern können. Probieren Sie es ruhig einmal aus.

Wie bereits angedeutet, können Sie auch die Position der Views beliebig ändern. Bewegen Sie den Mauszeiger hierzu einfach auf den Namensbereich des Views und draggen Sie den View an eine andere Position. Die Position, an der View nach dem Loslassen der Maustaste angeordnet werden würde, wird durch einen transparenten grauen Bereich markiert:

Sie können so einzelne Views zu mehreren Karteikarten zusammenfassen, so dass

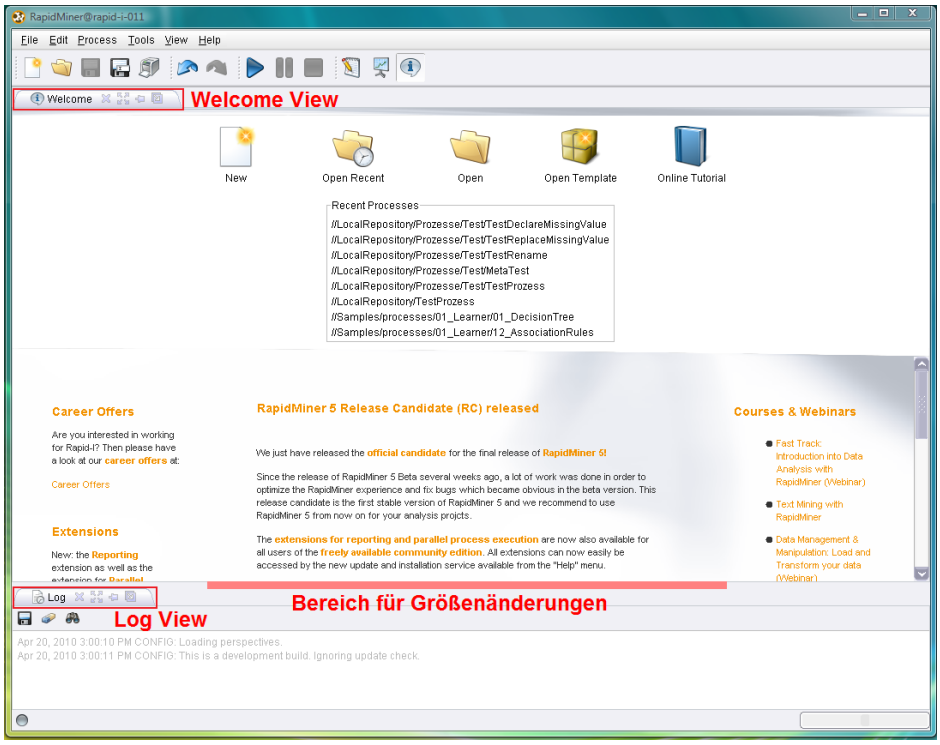

RapidMiner@rapid-i-011
File Edit Process Tools View Help
Welcome
Welcome View
New
Open Recent
Open
Open Template
Online Tutorial
Recent Processes
//LocalRepository/Prozesse/Test/TestDeclareMissingValue
//LocalRepository/Prozesse/Test/TestReplaceMissingValue
//LocalRepository/Prozesse/Test/TestRename
//LocalRepository/Prozesse/Test/MetaTest
//LocalRepository/Prozesse/Test/TestProzess
//LocalRepository/TestProzess
//Samples/processes/01_Learner/01_DecisionTree
//Samples/processes/01_Learner/12_AssociationRules
Career Offers
Are you interested in working for Rapid-I? Then please have a look at our career offers at:
Career Offers
Extensions
New: the Reporting extension as well as the
RapidMiner 5 Release Candidate (RC) released
We just have released the official candidate for the final release of RapidMiner 5!
Since the release of RapidMiner 5 Beta several weeks ago, a lot of work was done in order to optimize the RapidMiner experience and fix bugs which became obvious in the beta version. This release candidate is the first stable version of RapidMiner 5 and we recommend to use RapidMiner 5 from now on for your analysis projcts.
The extensions for reporting and parallel process execution are now also available for all users of the freely available community edition. All extensions can now easily be accessed by the new update and installation service available from the "Help" menu.
Courses & Webinars
Fast Track: Introduction into Data Analysis with RapidMiner (Webinar)
Text Mining with RapidMiner
Data Management & Manipulation: Load and Transform your data (Webinar)
Log
Log View
Bereich für Größenänderungen
Apr 20, 2010 3:00:10 PM CONFIG: Loading perspectives.
Apr 20, 2010 3:00:11 PM CONFIG: This is a development build. Ignoring update check.

Abbildung 2.5: Größenänderungen zwischen Views

stets nur einer sichtbar ist. Oder Sie ziehen den Log View von unten in den rechten Bereich, so dass die Teilung nun vertikal und nicht mehr horizontal verläuft. Sie können sogar einen View komplett abdocken und außerhalb des RapidMiner Fensters schieben. Möchten Sie einen View kurzzeitig vollständig sehen, so können Sie einen View maximieren und später wieder minimieren. Dies wird übrigens auch durchgeführt, wenn Sie einen Doppelklick auf den Namensbereich eines Views durchführen. Jeder View stellt Ihnen die folgenden Aktionen zur Verfügung:

Unter anderem die folgenden Aktionen sind für alle Views von RapidMiner möglich, weitere Aktionen sollten selbsterklärend sein:

1. *Close:* Schließt den View in der aktuellen Perspektive. Sie können den View erneut in der aktuellen oder einer anderen Perspektive öffnen mittels des Menüs „View“ – „Show View“.

2. *Maximize:* Maximiert den View in der aktuellen Perspektive. Kann auch

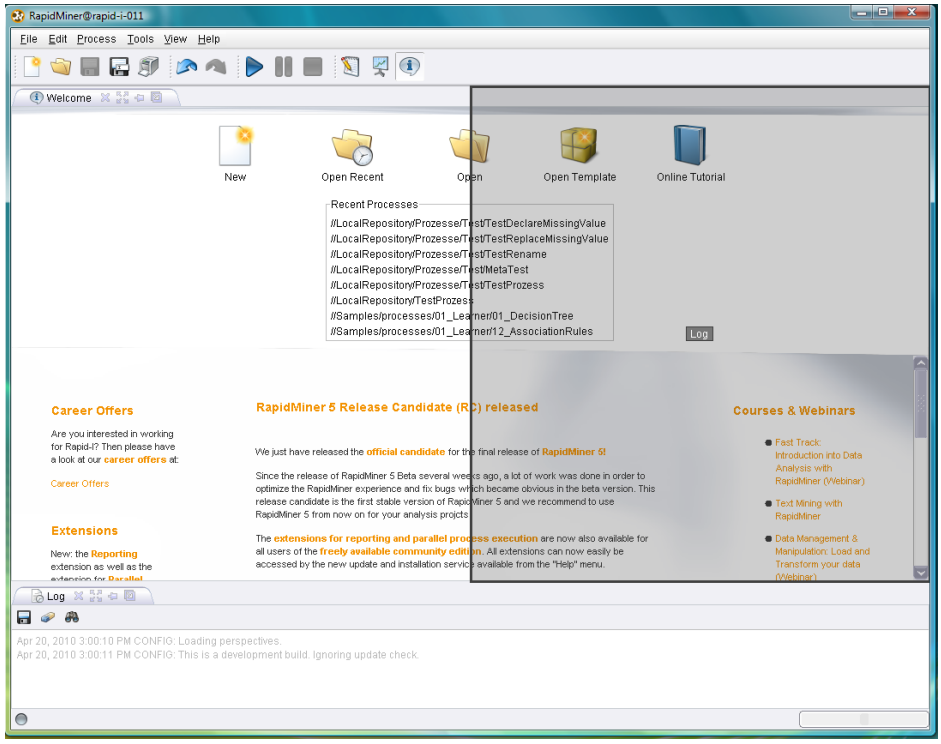

RapidMiner@rapid-i-011
File Edit Process Tools View Help
Welcome
New
Open Recent
Open
Open Template
Online Tutorial
Recent Processes
//LocalRepository/Prozesse/Test/TestDeclareMissingValue
//LocalRepository/Prozesse/Test/TestReplaceMissingValue
//LocalRepository/Prozesse/Test/TestRename
//LocalRepository/Prozesse/Test/MetaTest
//LocalRepository/Prozesse/Test/TestProzess
//LocalRepository/TestProzess
//Samples/processes/01_Learner/01_DecisionTree
//Samples/processes/01_Learner/12_AssociationRules
Log
Career Offers
Are you interested in working for Rapid-I? Then please have a look at our career offers at:
Career Offers
Extensions
New: the Reporting extension as well as the
RapidMiner 5 Release Candidate (RC) released
We just have released the official candidate for the final release of RapidMiner 5!
Since the release of RapidMiner 5 Beta several weeks ago, a lot of work was done in order to optimize the RapidMiner experience and fix bugs which became obvious in the beta version. This release candidate is the first stable version of RapidMiner 5 and we recommend to use RapidMiner 5 from now on for your analysis projcts.
The extensions for reporting and parallel process execution are now also available for all users of the freely available community edition. All extensions can now easily be accessed by the new update and installation service available from the "Help" menu.
Courses & Webinars
Fast Track: Introduction into Data Analysis with RapidMiner (Webinar)
Text Mining with RapidMiner
Data Management & Manipulation: Load and Transform your data (Webinar)
Log
Apr 20, 2010 3:00:10 PM CONFIG: Loading perspectives.
Apr 20, 2010 3:00:11 PM CONFIG: This is a development build. Ignoring update check.

Abbildung 2.6: Draggen des unteren Log-Views an die rechte Seite und Markierung der neuen Position.

Abbildung 2.7: Aktionen für Views

mittels Doppelklick auf den Namensbereich durchgeführt werden.

3. *Minimize:* Minimiert den View in der aktuellen Perspektive. Der View wird auf der linken Seite der Perspektive angezeigt und kann von dort aus wieder maximiert oder kurz betrachtet werden.

4. *Detach:* Löst den View aus der aktuellen Perspektive und stellt ihn innerhalb eines eigenen Fensters dar, welches beliebig verschoben werden kann.

Probieren Sie nun einfach ein wenig, die beiden Views auf verschiedene Arten anzuordnen. Es erfordert manchmal ein wenig Übung, die Views genau an der

gewünschten Stelle abzulegen. Es lohnt sich jedoch, ein wenig mit den Anordnungen zu experimentieren, denn je nach Bildschirmauflösung und persönlichen Präferenzen mögen andere Einstellungen Ihre Arbeit deutlich effizienter machen.

Manchmal löscht man versehentlich einen View oder verschiebt sich die Perspektive ungewollt in besonders ungünstige Varianten. In diesem Fall hilft das „View“ Menü weiter, denn neben der Möglichkeit, geschlossene Views mittels „Show View“ wieder zu öffnen, kann auch der ursprüngliche Zustand mittels „Restore Default Perspektive“ jederzeit wieder hergestellt werden.

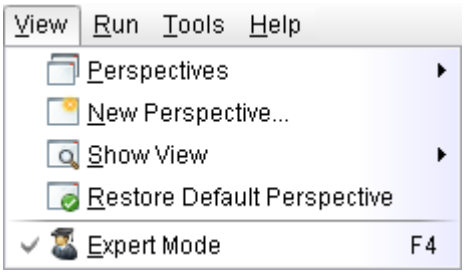


Abbildung 2.8: View Menü

Außerdem finden Sie hier noch die Möglichkeit, auch eigene Perspektiven unter einem frei wählbaren Namen abzuspeichern („New Perspective...“) sowie zwischen den gespeicherten und vordefinierten Perspektiven zu wechseln.

2.3 Design-Perspektive

Wie eingangs bereits erwähnt, finden Sie im rechten Bereich der Toolbar ein Icon für jede (vordefinierte) Perspektive:

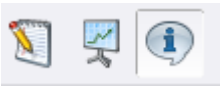


Abbildung 2.9: Toolbar Icons für Perspektiven

Die hier dargestellten Icons wechseln in die folgenden Perspektiven:

1. *Design-Perspektive:* Dies ist die zentrale Ansicht von RapidMiner in der alle Analyseprozesse erstellt und verwaltet werden.

2. *Result-Perspektive:* Wenn ein Prozess Ergebnisse in Form von Daten, Modellen o.ä. liefert, so wechselt RapidMiner in diese Ergebnisansicht, in der Sie wie gewohnt dank der Views auch mehrere Resultate gleichzeitig betrachten können.

3. *Welcome-Perspektive:* Die bereits oben beschriebene Willkommensansicht, mit der Sie RapidMiner nach dem Programmstart begrüßt.

Sie können mittels eines Klicks innerhalb der Toolbar in die gewünschte Perspektive wechseln oder alternativ mittels des Menüeintrags „View“ – „Perspectives“ gefolgt von der Auswahl der Zielperspektive. Schließlich fragt RapidMiner Sie auch automatisch, falls ein Wechsel in eine andere Perspektive sinnvoll scheint, beispielsweise zur Ergebnisansicht bei Beendigung eines Analyseprozesses.

Wechseln Sie nun mittels eines Klicks in der Toolbar in die Design-Perspektive. Sie wird im Rahmen dieses Kapitels ausführlich behandelt. Die Result-Perspektive wird dann Thema eines späteren Kapitels sein. Sie sollten nun den folgenden Bildschirm vor sich sehen:

Da es sich bei der Designansicht um die zentrale Arbeitsumgebung von RapidMiner handelt, werden wir im Folgenden alle Teile der Design-Perspektive einzeln besprechen und die grundlegenden Funktionalitäten der zugehörigen Views diskutieren.

2.3.1 Operators und Repositories View

In diesem Bereich finden sich zumindest in der Standardeinstellung zwei ausgesprochen zentrale Views, die im Folgenden beschrieben werden.

Operators View

Hier werden alle in RapidMiner verfügbaren Arbeitsschritte (Operatoren) in Gruppen präsentiert und stehen damit zum Einfügen in den aktuellen Prozess zur Verfügung. Sie können auf einfache Weise innerhalb der Gruppen navigieren und nach Herzenslust in den mitgelieferten Operatoren stöbern. Wenn RapidMiner mittels einer der erhältlichen Extensions erweitert wurde, so finden sich die zusätzlichen Operatoren ebenfalls an dieser Stelle.

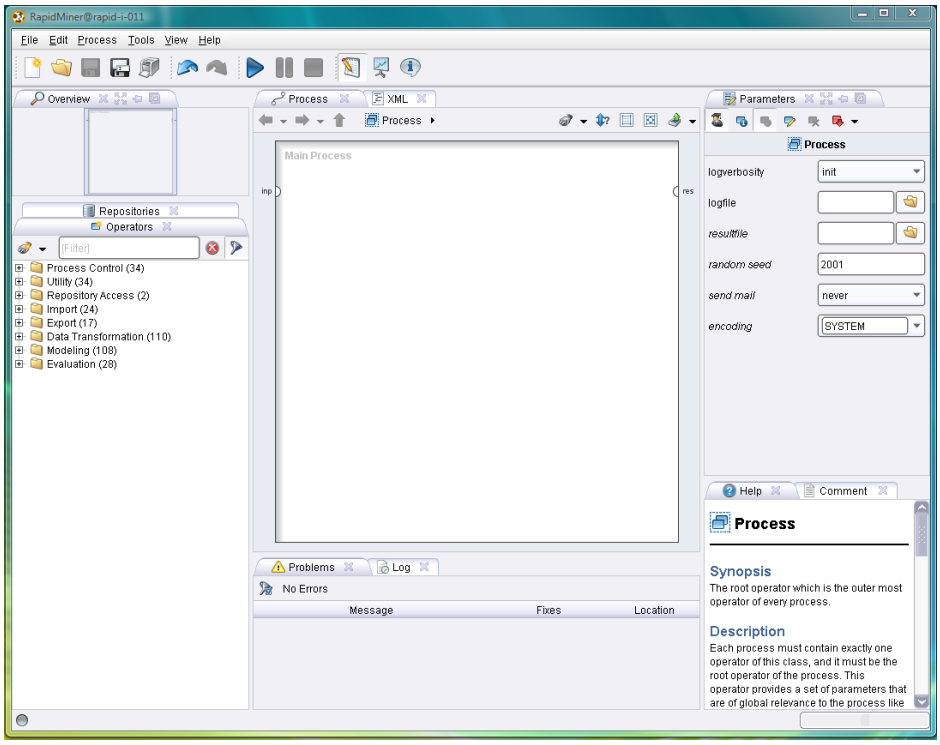


Abbildung 2.10: Design-Perspektive von RapidMiner

Ohne Extensions finden Sie zumindest die folgenden Gruppen von Operatoren in der Baumstruktur:

- *Process Control:* Operatoren wie Loops oder Bedingte Verzweigungen, welche den Prozessfluss steuern können.
- *Utility:* Hilfsoperatoren, die neben dem Operator „Subprocess“ zur Gruppierung von Unterprozessen auch die wichtigen Makro-Operatoren sowie die Operatoren zum Logging beinhalten.
- *Repository Access:* Enthält die beiden Operatoren für lesenden und schreibenden Zugriff in Repositories.
- *Import:* Enthält eine Vielzahl von Operatoren, um Daten und Objekte aus externen Formaten lesen zu können wie Dateien, Datenbanken usw.
- *Export:* Enthält eine Vielzahl von Operatoren, um Daten und Objekte in

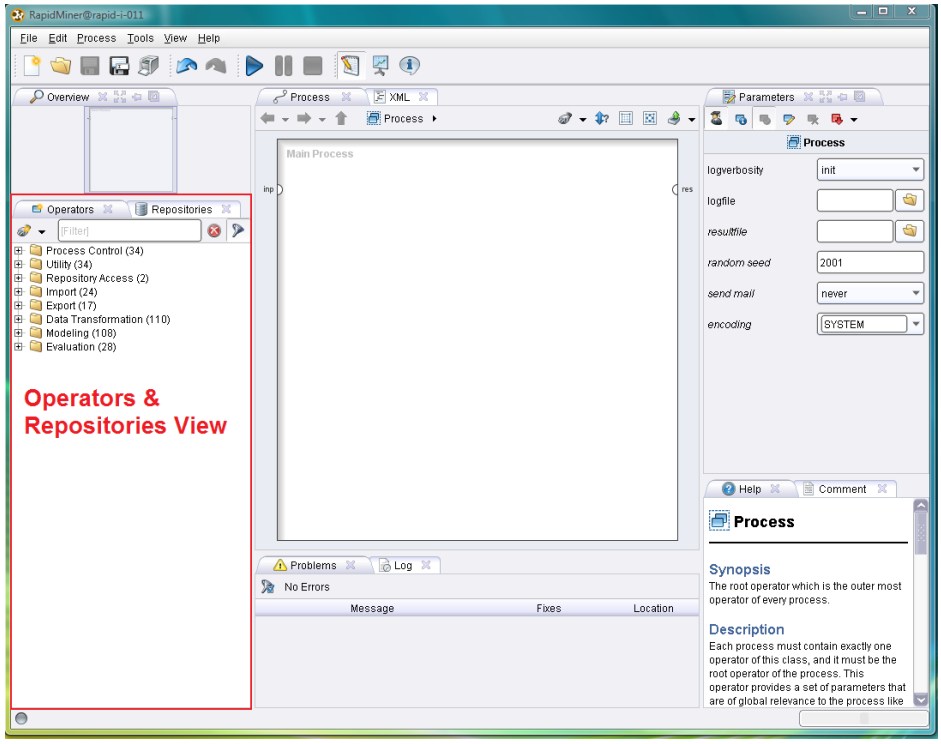


Abbildung 2.11: Design-Operatoren von RapidMiner

externe Formate schreiben zu können wie Dateien, Datenbanken usw.

- *Data Transformation:* Die gemessen an Umfang und Bedeutung in der Analyse wohl wichtigste Gruppe. Hier befinden sich alle Operatoren um sowohl Daten als auch Metadaten transformieren zu können.

- *Modeling:* Enthält die eigentlichen Data Mining Verfahren wie Klassifikationsverfahren, Regressionsverfahren, Clustering, Gewichtungen, Verfahren für Assoziationsregeln, Korrelations- und Ähnlichkeitsanalysen sowie Operatoren, um die generierten Modelle auf neue Datensätze anzuwenden.

- *Evaluation:* Operatoren, mit deren Hilfe man die Güte einer Modellierung berechnen und damit für neue Daten abschätzen kann wie Kreuzvalidierungen, Bootstrapping usw.

Sie können neue Operatoren innerhalb des Operators View einfach auswählen

und mittels Drag&Drop an der gewünschten Stelle im Prozess hinzufügen. Sie können dabei wählen, ob neue Operatoren direkt möglichst passend auf Basis der vorliegenden Metadaten-Informationen mit bereits bestehenden Operatoren verbunden werden oder nicht. Wählen Sie dazu einfach das Stecker-Symbol links in der Toolbar des Views und definieren Sie, ob eingehende und / oder ausgehende Verbindungen automatisch erzeugt werden sollen. Andernfalls müssen Sie den Operator selbständig verbinden.

Abbildung 2.12: Aktionen und Filter für den Operators View

Um Ihnen die Arbeit möglichst zu erleichtern, unterstützt der Operators View zudem noch einen Filter, welcher verwendet werden kann, um nach Bestandteilen des Operatornamens beziehungsweise dem vollständigen Operatornamen zu suchen. Geben Sie einfach den Suchbegriff in das Filterfeld ein. Sobald insgesamt weniger als 10 Suchtreffer existieren, wird der Baum so aufgeklappt, dass alle Suchtreffer sichtbar sind. So brauchen Sie nicht jedes Mal durch die vollständige Hierarchie zu navigieren. Ein Klick auf das rote Kreuz neben dem Suchfeld löscht die aktuelle Eingabe und klappt den Baum wieder zusammen.

Tipp: Profis werden mit der Zeit die Namen der benötigten Operatoren immer häufiger kennen. Das Suchfeld unterstützt neben der Suche nach dem (vollständigen) Namen auch eine Suche auf Basis der Anfangsbuchstaben (sogenannte Camel-Case-Search). Probieren Sie einfach mal „REx“ für „Read Excel“ oder „DN“ für „Date to Nominal“ und „Date to Numerical“ – dies beschleunigt die Suche nochmals enorm.

Repositories View

Das Repository ist ein zentraler Bestandteil von RapidMiner, der mit Version 5 Einzug gehalten hat. Es dient der Verwaltung und Strukturierung Ihrer Analyseprozesse in Projekte und zugleich auch als Quelle sowohl von Daten als auch der zugehörigen Metadaten. Die Verwendung des Repositories erläutern wir Ihnen ausführlich in den nächsten Kapiteln, daher belassen wir es an dieser Stelle lediglich bei dem folgenden.

Hinweis: Da ein Großteil der Unterstützungen von RapidMiner für das Prozess-Design von Metadaten Gebrauch macht, empfehlen wir Ihnen dringend die Verwendung des Repositories, da andernfalls, beispielsweise bei unmittelbaren Lesen von Daten aus Dateien oder Datenbanken, die Metadaten nicht zur Verfügung stehen und so zahlreiche Unterstützungen nicht angeboten werden.

2.3.2 Process View

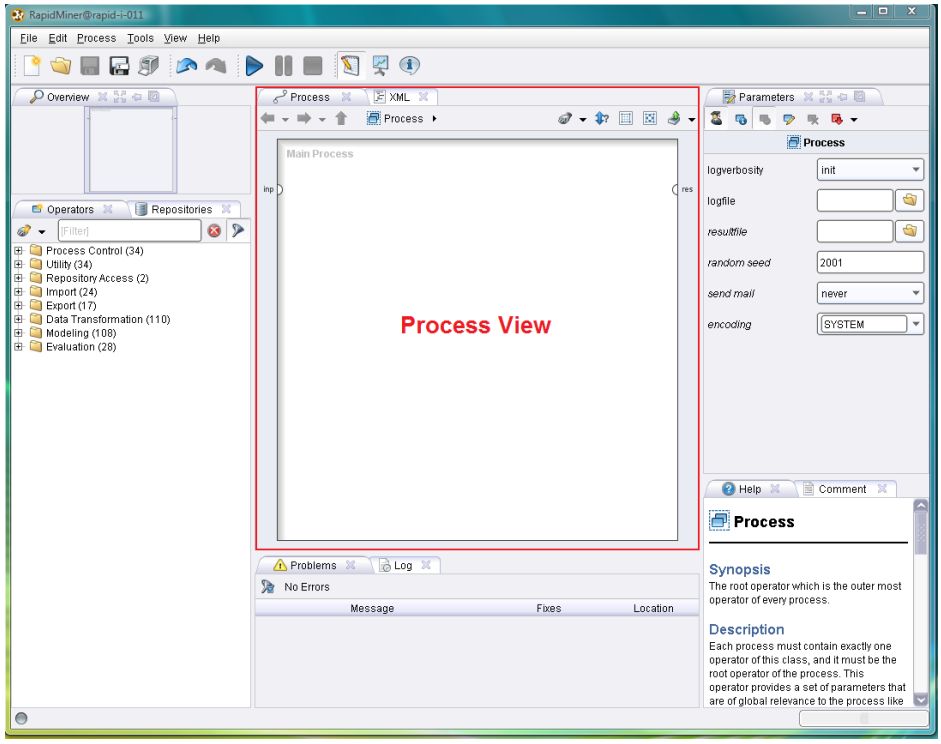


Abbildung 2.13: Im Process View werden die Bausteine von RapidMiner, die sogenannten Operatoren, miteinander verbunden

Der Process View zeigt die einzelnen Schritte innerhalb des Analyseprozesses sowie deren Verbindungen untereinander. Neue Schritte können auf mehrere Weisen dem aktuellen Prozess hinzugefügt werden. Verbindungen zwischen diesen Schritten können definiert und wieder gelöst werden. Schließlich ist es sogar möglich, sogar die Reihenfolge der Schritte in dieser Ansicht zu definieren. Aber eins nach dem anderen.

2.3.3 Operatoren und Prozesse

Die grundlegende Arbeitsweise mit RapidMiner besteht in der Definition von Analyseprozessen durch die Angabe einer Abfolge von einzelnen Arbeitsschritten. In RapidMiner heißen diese Prozessbausteine **Operatoren**. Ein Operator ist durch mehrere Dinge definiert:

- die Beschreibung der erwarteten Eingaben,
- die Beschreibung der gelieferten Ausgaben,
- die Aktion, die der Operator auf den Eingaben ausführt und welche schließlich die Ausgabe berechnet,
- eine Menge von Parametern, welche die durchgeführte Aktion steuern können.

Die Ein- und Ausgaben von Operatoren werden über **Ports** generiert beziehungsweise konsumiert. Wir werden sehen, dass in RapidMiner ein Operator durch einen Baustein in der folgenden Form dargestellt wird:

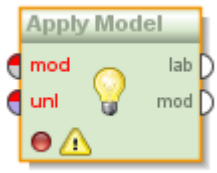


Abbildung 2.14: Ein Operator kann über seine Input-Ports (links) und Output-Ports (rechts) verbunden werden.

Ein solcher Operator kann beispielsweise Daten aus dem Repository, einer Datenbank oder aus Dateien einlesen. In diesem Fall hätte er keine Input-Ports, wohl aber Parameter, der zumindest den Ort der Daten spezifiziert. Andere Operatoren transformieren ihre Eingaben und liefern ein Objekt des gleichen Typs zurück. Operatoren, die Daten transformieren, gehören in diese Gruppe. Und wieder andere Operatoren konsumieren ihre Eingabe und verwandeln diese in ein vollständig neues Objekt: viele Data Mining Verfahren gehören hierzu und liefern beispielsweise ein Modell für die gegebenen Input-Daten.

Die Farbe der Ports gibt an, mit welchem Eingabetyp ein Port versorgt werden muss. Ein bläulicher Farbton beispielsweise zeigt an, dass eine Beispielmenge

(Example Set) verlangt wird. Ist die obere Hälfte und der Name des Ports rot eingefärbt, so deutet dies auf ein Problem hin. Für den Operator oben ist dieses Problem leicht zu sehen: er ist nicht verbunden und die Input-Ports benötigen noch eine Verbindung zu einer passenden Quelle.

Weiße Output-Ports liegen dann vor, wenn das Resultat unklar ist beziehungsweise in der derzeitigen Konfiguration (noch) nicht geliefert werden kann. Sobald alle notwendigen Konfigurationen abgeschlossen wurden, d.h. alle notwendigen Parameter definiert und alle notwendigen Input-Ports verbunden, so färben sich die Output-Ports gemäß ihres Typs ebenfalls ein.

Abbildung 2.15: Statusanzeigen von Operatoren

Aber nicht nur die Ports, sondern auch der komplette Operator kann seinen Zustand mittels verschiedener Statusanzeigen visualisieren. Diese sind von links nach rechts gegeben durch:

- *Statusampel:* Zeigt an, ob ein Problem vorliegt wie noch nicht eingestellte Parameter oder unverbundene Input-Ports (rot), ob die Konfiguration prinzipiell abgeschlossen ist aber der Operator seitdem noch nicht ausgeführt wurde (gelb) oder ob alles in Ordnung ist und der Operator auch bereits erfolgreich durchgeführt wurde (grün).

- *Warndreieck:* Zeigt an, wenn für diesen Operator Statusmeldungen vorliegen.

- *Breakpoint:* Zeigt an, ob die Prozessausführung vor oder nach diesem Operator angehalten werden soll, um dem Analysten die Gelegenheit zu geben, Zwischenergebnisse zu inspizieren.

- *Kommentar:* Wenn ein Kommentar zu diesem Operator eingegeben wurde, so wird dies mittels dieses Icons angezeigt.

- *Subprozess:* Dies ist eine sehr wichtige Anzeige, da manche Operatoren über einen oder mehrere Unterprozesse verfügen. Ob ein solcher Unterprozess existiert, wird mittels dieses Zeichens angezeigt. Sie können einen Dop-

pelklick auf den betreffenden Operator ausführen, um in die Unterprozesse abzusteigen.

Werden mehrere Operatoren miteinander verbunden, so sprechen wir von einem **Analyseprozess** oder kurz **Prozess**. Eine solche Abfolge von Schritten kann beispielsweise einen Datensatz einladen, die Daten transformieren, ein Modell berechnen und das Modell auf einen anderen Datensatz anwenden. In RapidMiner kann ein solcher Prozess dann wie folgt aussehen:

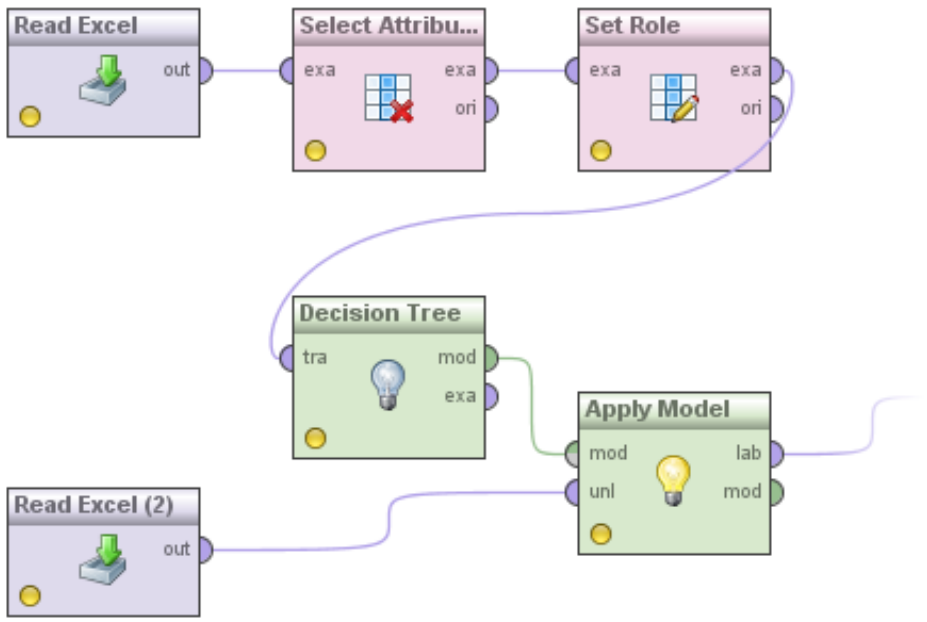


Abbildung 2.16: Ein Analyseprozess bestehend aus mehreren Operatoren. Die Farbkodierung der Datenflüsse gibt den Typ des weitergegebenen Objekts wieder.

In RapidMiner können solche Prozesse leicht mehrere hundert Operatoren groß werden und sich über mehrere Ebenen bzw. Subprozesse hinziehen. Die stets im Hintergrund durchgeführten Prozessprüfungen wie auch die unten dargestellten Hilfsmittel zur Prozessnavigation stellen sicher, dass Sie den Überblick nicht verlieren und auch für komplexere Aufgaben korrekte Prozesse definieren.

Einfügen von Operatoren

Sie können auf verschiedene Weisen neue Operatoren in den Prozess einfügen. Die verschiedenen Möglichkeiten sind im Einzelnen:

- via Drag&Drop aus dem Operators View wie oben beschrieben,
- via Doppelklick auf einen Operator im Operators View,
- via Dialog, welcher mittels des ersten Icons in der Toolbar des Process Views geöffnet wird,
- via Dialog, welcher mittels des Menüeintrags „Edit“ – „New Operator. . . “ geöffnet wird (CTRL-I),
- via Kontextmenü in einem freien Bereich der weißen Prozessfläche und dort mittels des Untermenüs „New Operator“ und durch Auswahl eines Operators.

Abbildung 2.17: Aktionen im Process View

In jedem Fall gilt, dass neue Operatoren abhängig von der Einstellung im Operators View entweder automatisch mit passenden Operatoren verbunden werden oder dass die Verbindungen manuell durch den Anwender nun erfolgen bzw. korrigiert werden muss.

Verbinden von Operatoren

Nachdem Sie neue Operatoren eingefügt haben, können Sie die eingefügten Operatoren miteinander verbinden. Dazu stehen Ihnen prinzipiell drei Wege offen, die im Folgenden beschrieben werden.

Verbindungen 1: Automatisch beim Einfügen

Sollten Sie im Operators View die Option zum automatischen Verbinden unter dem Stecker-Symbol aktiviert haben, so wird RapidMiner nach dem Einfügen

versuchen, den Operator mit passenden Output-Ports zu verbinden. Sollte beispielsweise der neue Operator einen Input-Port besitzen, der eine Beispielmenge verlangt, so wird RapidMiner versuchen einen Operator zu finden, der eine solche Beispielmenge bereits produzieren könnte. Wenn es nur eine Option gibt, so ist diese Wahl eindeutig und der Operator wird verbunden. Wenn es jedoch mehrere Optionen gibt, dann versucht RapidMiner diejenige Option zu wählen, die sich am nächsten links oberhalb von der aktuellen Mausposition befindet. Der zugehörige Operator wird mit einem Rahmen und einem Schatten markiert. Auf diese Weise können Sie bereits während des Einfügens die Weichen für eine korrekte Verbindung stellen.

Tipp: Es ist empfehlenswert, die Option zum automatischen Verbinden zumindest für die Input-Ports zu aktivieren. Auch wenn von Zeit zu Zeit der Verbindungsalgorithmus auf Basis der Metadaten eine falsche Verbindung erzeugt, so sparen Sie sich viel Arbeit für all die Fälle, in denen die korrekte Verbindung automatisch erkannt wird.

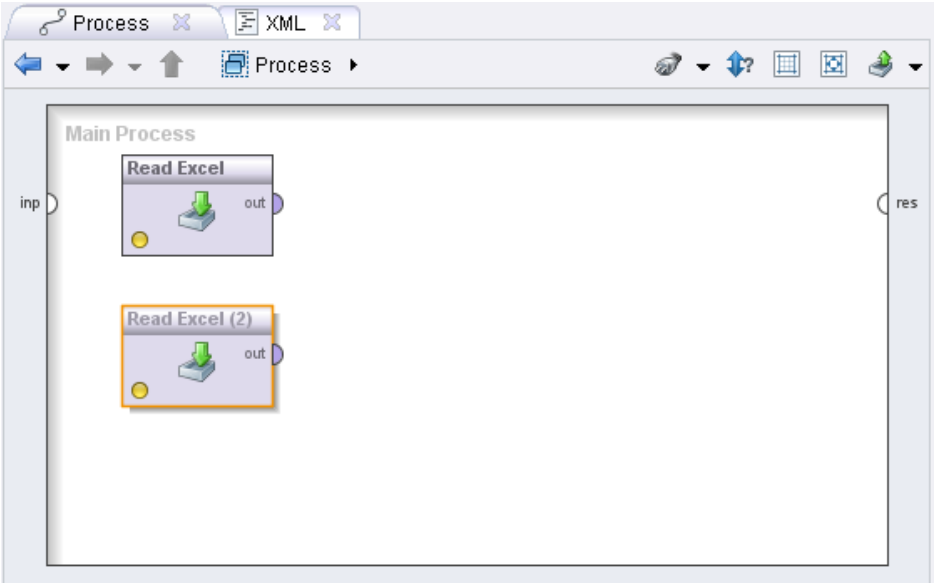


Abbildung 2.18: Der zweite Operator wird während des Drag-Vorgangs markiert (Rahmen plus Schatten) und wird bevorzugt mit dem neuen Operator verbunden, wenn dieser jetzt fallen gelassen wird und sofern dieser eine Beispielmenge erwartet.

Verbindungen 2: Manuell

Sie können die Operatoren auch manuell miteinander verbinden und tatsächlich ist dieses bei komplexeren Prozessen auch notwendig. Klicken Sie hierzu auf einen Output-Port. Sie zeichnen nun eine orangefarbene Linie. Klicken Sie auf einen Input-Port um den gewählten Output-Port mit diesem Input-Port zu verbinden. Um den Vorgang abzubrechen, halten Sie die Maus still und klicken Sie mit der rechten Maustaste. Die orangefarbene Linie verschwindet und Sie können wie gewohnt weiterarbeiten.

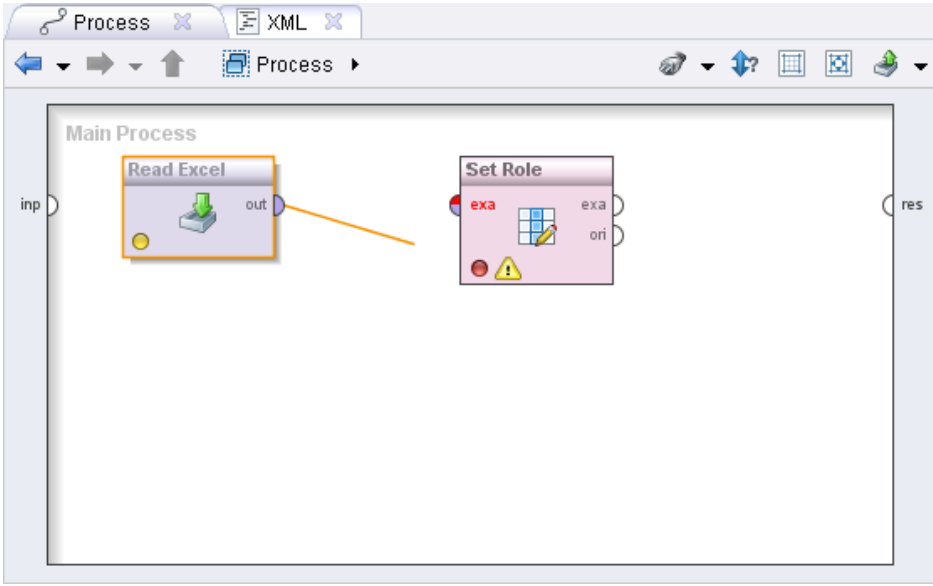


Abbildung 2.19: Klick auf einen Output-Port zum Verbinden, Rechtsklick zum Abbrechen.

Verbindungen 3: Vollautomatisch

Manchmal befinden sich schon zahlreiche Operatoren in einem (Sub-)Prozess und sind noch nicht verbunden. In einem solchen Fall kann die Option „Auto-Wire“ bzw. „Re-Wire“ gute Dienste leisten, welche sich hinter dem Stecker-Symbol im Process View direkt neben dem Icon für das Öffnen des Dialogs für einen neuen Operator verbergen. Dies funktioniert insbesondere dann gut, wenn bereits bei der Erstellung des Prozesses relativ sequentiell vorgegangen wurde und die Operatoren ordentlich hintereinander aufgereiht wurden, d.h. immer der vorherige Operator beim Einfügen durch Rahmen und Schatten markiert war. Eine manuelle Prüfung nach der vollautomatischen Verbindung ist allerdings in jedem Fall sinnvoll, da es gerade bei komplexeren Prozessen hierbei zu nicht beabsichtigten

Verbindungen kommen kann.

Auswählen von Operatoren

Zum Editieren von Parametern müssen Sie einen einzelnen Operator auswählen. Sie erkennen den aktuell ausgewählten Operator an seinem orangefarbenen Rahmen zusammen mit einem Schatten.

Wenn Sie eine Aktion für mehrere Operatoren gleichzeitig durchführen wollen, beispielsweise Bewegen oder Löschen, so wählen Sie bitte alle gewünschten Operatoren aus, indem Sie einen Rahmen um diese ziehen.

Um einzelne Operatoren der aktuellen Auswahl hinzu zu fügen beziehungsweise um einzelne Operatoren aus der aktuellen Auswahl auszuschließen, halten Sie bitte die Taste STRG gedrückt, während Sie auf die gewünschten Operatoren klicken beziehungsweise weitere Operatoren mittels Ziehen eines Rahmens hinzufügen.

Bewegen von Operatoren

Wählen Sie einen oder mehrere Operatoren wie oben beschrieben aus. Bewegen Sie nun den Mauszeiger auf einen der ausgewählten Operatoren und ziehen Sie die Maus bei gedrückter Taste. Alle ausgewählten Operatoren werden nun gemäß der Mausbewegung an eine neue Stelle bewegt.

Falls Sie im Zuge dieser Bewegung den Rand der weißen Fläche erreichen, so wird diese automatisch entsprechend vergrößert. Sollten Sie an den Rand des sichtbaren Bereichs kommen, so wird dieser ebenfalls automatisch direkt mit verschoben.

Löschen von Operatoren

Wählen Sie einen oder mehrere Operatoren wie oben beschrieben aus. Sie können die ausgewählten Operatoren nun Löschen mittels

- Drücken der Taste ENTFERNEN,
- Auswahl der Aktion „Delete“ im Kontextmenü einer der ausgewählten Operatoren,

- des Menüeintrags „Edit“ – „Delete“.

Löschen von Verbindungen

Verbindungen können durch Klicken auf einen der beiden Ports bei gleichzeitigem Drücken der Taste ALT gelöscht werden. Alternativ können Sie eine Verbindung auch mittels den Kontextmenüs der betroffenden Ports löschen.

Navigieren im Prozess

Betrachten wir noch einmal die Toolbar für den Process View, so stellen wir fest, dass wir bisher lediglich von den linken beiden Aktionen Gebrauch gemacht haben. Die folgenden vier Elemente, nämlich den Pfeil nach Links, den Pfeil nach rechts, den Pfeil nach oben und die Navigationsleiste (Breadcrumb) diskutieren wir in diesem Abschnitt.

Abbildung 2.20: Aktionen im Process View

Die Aktionen im Einzelnen:

1. *Pfeil nach links:* Kehrt zur letzten Editierstelle zurück analog zur Navigation, welche aus Internetbrowsern bekannt ist. Einzelne Schritte können mittels des Ausklappmenüs auch übersprungen werden.

2. *Pfeil nach rechts:* Wieder zu in der Historie weiter vorne liegenden Editierstellen nach vorne gehen analog zur Navigation, welche aus Internetbrowsern bekannt ist. Einzelne Schritte können mittels des Ausklappmenüs auch übersprungen werden.

3. *Pfeil nach oben:* Aus dem aktuellen Subprozess wieder in den überge“-ordneten Prozess zurückkehren.

4. *Navigationsleiste:* Die Navigationsleiste zeigt den Weg vom Hauptprozess über alle gegangenen Ebenen in den aktuellen Subprozess an. Ein Klick auf

einen der Operatoren zeigt den betreffenden Prozess. Mittels der kleinen Pfeile nach rechts kann weiter abwärts navigiert werden.

Um also in einen Unterprozess hinab zu steigen, ist ein Doppelklick auf einen Operator mit dem Subprozess-Icon unten rechts nötig. Um wieder eine Ebene nach oben zu gehen, kann mittels des Pfeils nach oben navigiert werden. Den aktuellen Pfad zeigt die Navigationsleiste, die alternativ auch zur Navigation in beide Richtungen verwendet werden kann.

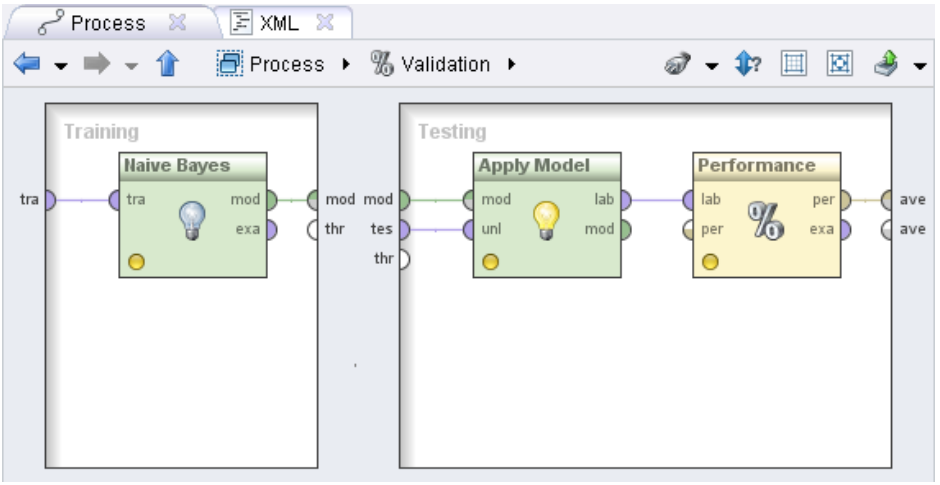


Abbildung 2.21: Ein Subprozess namens „Validation“, der mittels Pfeil nach oben oder über die Navigationsleiste wieder verlassen werden kann.

Definition der Ausführungsreihenfolge

In fast allen Fällen gelingt es RapidMiner automatisch, die korrekte Ausführungsreihenfolge der Operatoren zu bestimmen. RapidMiner verwendet hierzu die Verbindungsinformationen und die Tatsache, dass ein Operator, dessen Ergebnis von einem anderen verwendet werden soll, natürlich vor diesem ausgeführt werden muss.

Es gibt jedoch Fälle, bei denen die Reihenfolge nicht automatisch festgelegt werden kann wie bei vollständig parallelen Teilprozessen oder bei der die automatische Reihenfolge nicht korrekt ist, beispielsweise weil ein Makro zunächst berechnet werden muss, bevor man es als Parameter in einem späteren Operator anwenden kann. Aber auch andere Gründe wie beispielsweise eine effizientere

Datenbehandlung oder eine exakt gewünschte Reihenfolge zur Ausführung beispielsweise für Reporting spielen häufig eine große Rolle.

Zu diesem Zweck bietet RapidMiner eine elegante Methode, die Reihenfolge der Operatoren anzuzeigen und die Ausführungsreihenfolge sogar bequem zu editieren. Hierzu klicken Sie bitte auf den Doppelpfeil nach oben und unten mit dem Fragezeichen in der Toolbar des Process Views und wechseln Sie so in die Ansicht zur Reihenfolgendefinition. Nun wird statt des Icons für jeden Operator die Nummer seiner Ausführung dargestellt. Der transparente orangefarbene Strang verbindet die Operatoren in dieser Reihenfolge miteinander, wie in Abbildung 2.22 zu sehen ist.

Um eine solche Reihenfolge zu ändern, kann an jeder beliebigen Stelle auf einen Operator geklickt werden. Der Pfad bis zu diesem Operator nun nicht geändert werden, aber die Wahl eines Operators, der nach dem gewählten kommt durch einen weiteren Klick, versucht die Reihenfolge so zu ändern, dass der zweite Operator möglichst schnell nach dem ersten ausgeführt wird. Während Sie die Maus über die übrigen Operatoren bewegen, sehen Sie die aktuelle Wahl in orange bis zu diesem Operator und in grau ab diesem. Eine unmögliche Wahl wird durch eine rote Zahl symbolisiert. Sie können eine aktuelle Auswahl mittels eines Rechtsklicks abbrechen. Mit nur wenigen Klicks können Sie, wie in Abbildung 2.23 abgebildet, auf diese Weise die Reihenfolge des oben beschriebenen Prozesses in die Folgende ändern.

2.3.4 Weitere Optionen des Process Views

Nachdem wir fast alle Optionen dieses zentralen Elements der Design-Perspektive von RapidMiner diskutiert haben, beschreiben wir nun noch die übrigen Aktionen in der Toolbar, die in Abbildung 2.24 zu sehen ist, sowie weitere Möglichkeiten des Process Views.

Die rechten drei Icons in der Toolbar des Process Views führen die folgenden Aktionen aus:

1. *Automatische Anordnung:* Ordnet alle Operatoren des derzeitigen Prozesses neu an gemäß der Verbindungen und der aktuellen Ausführungsreihenfolge.

2. *Automatische Größe:* Ändert die Größe der weißen Arbeitsfläche derart, dass alle derzeitig positionierten Operatoren gerade ausreichend Platz ha-

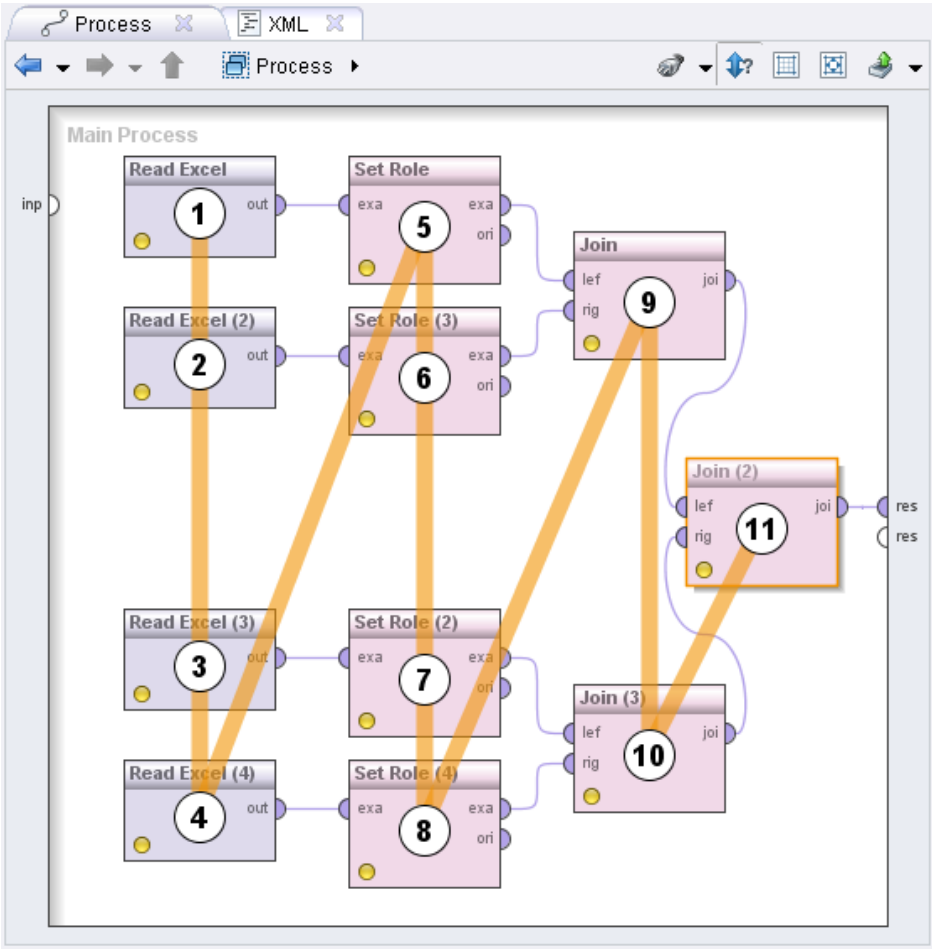


Abbildung 2.22: Darstellung der Ausführungsreihenfolge. Diese Reihenfolge ist jedoch ungünstig, da so mehr Datensätze gleichzeitig behandelt werden müssen.

ben. Dies ist insbesondere praktisch zur automatischen Verkleinerung (Größenoptimierung).

3. *Export:* Die aktuelle Prozessansicht kann sowohl gedruckt als auch nach PDF und in andere Formate exportiert werden.

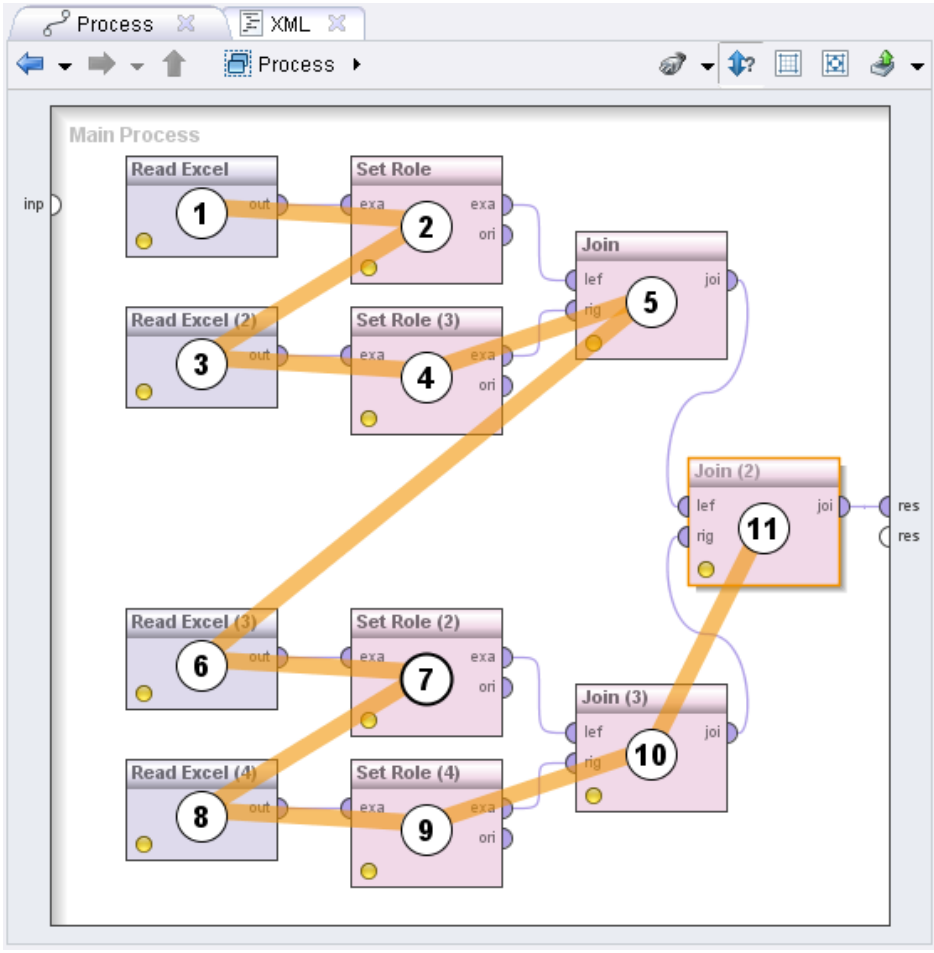


Abbildung 2.23: Neue Reihenfolge nach einigen Änderungen.

Abbildung 2.24: Aktionen im Process View

2.3.5 Parameters View

Abbildung 2.25 zeigt die Parameters View von RapidMiner.

Zahlreiche Operatoren benötigen für eine korrekte Funktionsweise die Angabe ei-

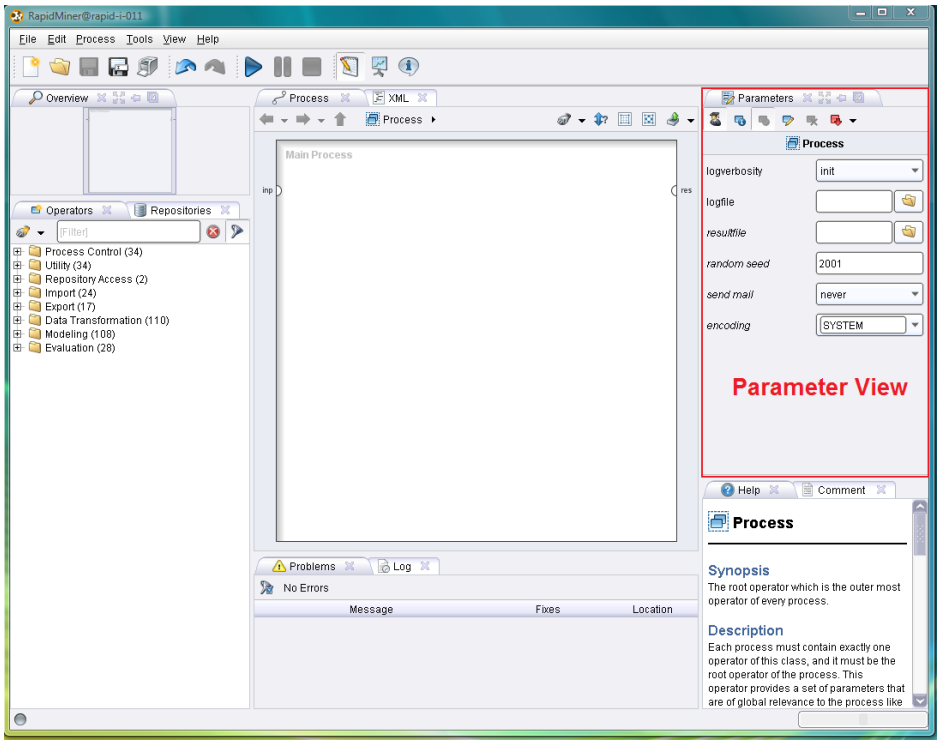


Abbildung 2.25: Parameter des aktuell selektierten Operators werden im Parameter View eingestellt.

nes oder mehrerer Parameter, beispielsweise benötigen Operatoren, die Daten aus Dateien lesen, die Angabe des Dateipfads. Noch viel häufiger jedoch kommt es vor, dass Parameter zwar nicht unbedingt erforderlich sind, jedoch die Ausführung des Operators durch die Angabe bestimmter Parameterwerte gesteuert und im Falle einer Modellierung häufig auch optimiert werden kann.

Nachdem ein Operator, der Parameter anbietet, im Process View selektiert wurde, werden seine Parameter im Parameter View angezeigt. Der View besitzt wie die anderen Views auch eine eigene Toolbar, die im Folgenden beschrieben wird. Unter der Toolbar finden sich Icon und Name des aktuell ausgewählten Operators gefolgt von den eigentlichen Parametern. Dabei bedeutet eine fette Schriftart, dass der Parameter unbedingt durch den Analysten definiert werden muss und keinen Default-Wert aufweist. Eine kursive Schrift bedeutet, dass der Parameter als Expertenparameter eingestuft wird und von Anfängern der Datenanalyse nicht unbedingt geändert werden sollte.

Bitte beachten Sie, dass manche Parameter erst dann angezeigt werden, wenn andere Parameter einen bestimmten Wert aufweisen. So kann beispielsweise für den Operator „Sampling“ nur dann eine absolute Anzahl gewünschter Beispiele angegeben werden, wenn als Typ des Samplings „absolute“ gewählt wurde.

Die Aktionen der Toolbar beziehen sich – genau wie die Parameter – auf den aktuell ausgewählten Operator. Im Einzelnen sind dies:

1. *Operator Info:* Anzeige einiger grundlegender Informationen zu diesem Operator wie erwartete Eingaben oder eine Beschreibung. Dieser Dialog wird auch durch Drücken von F1 nach Selektion, über das Kontextmenü im Process View sowie über den Menüeintrag „Edit“ – „Show Operator Info...“ angezeigt.

2. *Enable / Disable:* Operatoren können (vorübergehen) deaktiviert werden. Dabei werden ihre Verbindungen gelöst und sie werden nicht länger ausgeführt. Deaktivierte Operatoren werden grau dargestellt. Operatoren können auch innerhalb ihres Kontextmenüs im Process View sowie über den Menüeintrag „Edit“ – „Enable Operator“ (de-)aktiviert werden.

3. *Rename:* Eine der Möglichkeiten, einen Operator umzubenennen. Weitere Möglichkeiten sind das Drücken von F2 nach Selektion, die Auswahl „Rename“ im Kontextmenü des Operators im Process View sowie der Menüeintrag „Edit“ – „Rename“.

4. *Delete:* Eine der Möglichkeiten, einen Operator zu löschen. Weitere Möglichkeiten sind das Drücken von ENTFERNEN nach Selektion, die Auswahl „Delete“ im Kontextmenü des Operators im Process View sowie der Menüeintrag „Edit“ – „Delete“.

5. *Toggle Breakpoints:* Hier können Breakpoints sowohl vor als auch nach der Ausführung des Operators gesetzt werden, an denen die Prozessausführung stoppt und Zwischenergebnisse inspiziert werden können. Diese Möglichkeit besteht auch im Kontextmenü des Operators im Process View sowie im „Edit“-Menü. Ein Breakpoint nach Ausführung des Operators kann auch durch F7 aktiviert und deaktiviert werden.

6. *Flag as Dirty:* Setzt den Zustand des Operators wieder so ein, dass er bei wiederholter Prozessausführung in jedem Fall durchgeführt wird.

7. *Expert Mode:* Das Icon ganz rechts schaltet zwischen dem Expertenmodus und dem Anfängermodus um. Nur im Expertenmodus werden alle Parameter gezeigt, im Anfängermodus werden die als Expertenparameter eingestuften Parameter nicht dargestellt.

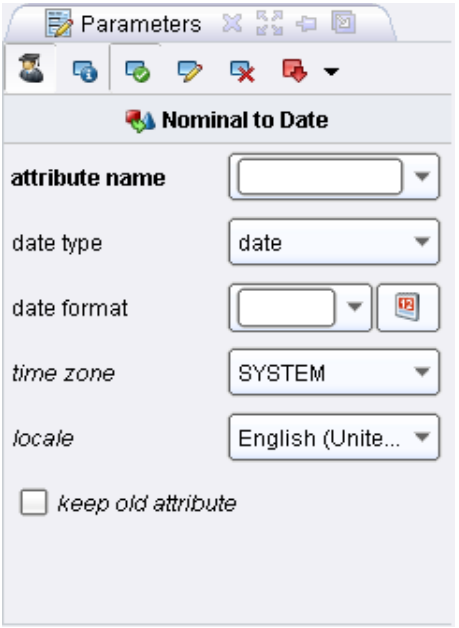


Abbildung 2.26: Die Parameter des Operator „Nominal to Date“.

2.3.6 Help und Comment View

Operators View

Jedes Mal, wenn Sie einen Operator im Operators View oder im Process View auswählen, zeigt das Hilfefenster im Rahmen des Help View eine Beschreibung dieses Operators an. Diese Beschreibungen umfassen

- eine kurze Synopsis, die die Funktion des Operators in einem oder wenigen Sätzen zusammenfasst,
- eine ausführliche Beschreibung der Funktionsweise des Operators,
- eine Auflistung aller Parameter inklusive einer kurzen Beschreibung des

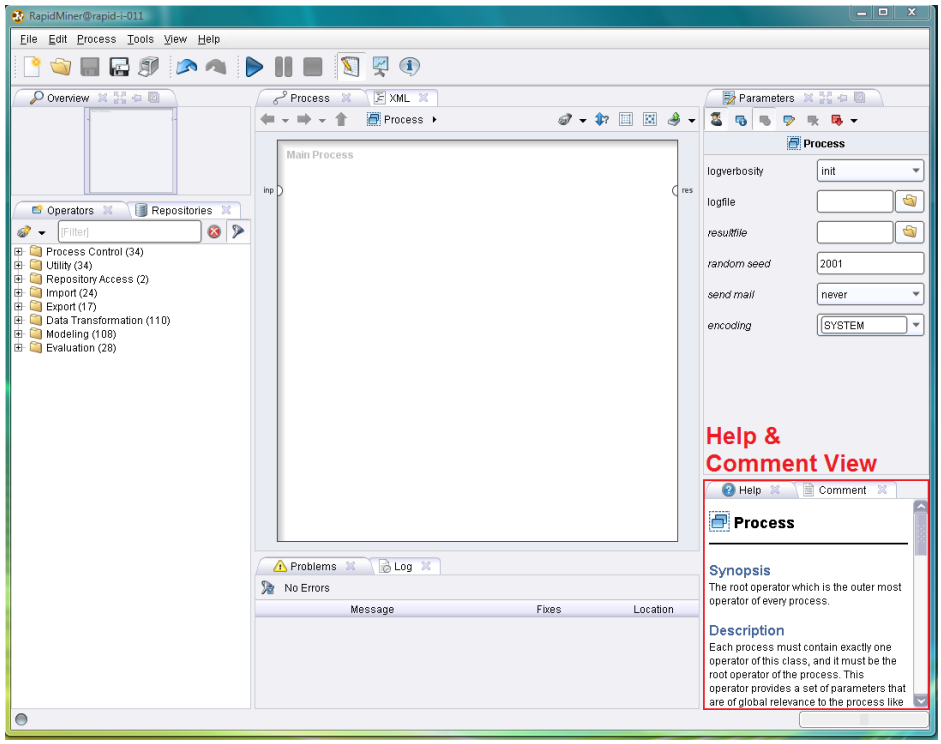


Abbildung 2.27: Sowohl zu aktuell ausgewählten Operatoren im Operators View als auch zu denen aus dem Process View werden Hilfstexte angezeigt.

Parameters, dem Default-Wert (falls vorhanden), der Angabe, ob es sich bei diesem Parameter um einen Expertenparameter handelt sowie einer Angabe von Parameterabhängigkeiten.

Comment View

Der Comment View ist im Gegensatz zur Hilfe nicht vordefinierten Beschreibungen sondern vielmehr Ihren eigenen Kommentaren zu einzelnen Schritten des Prozesses gewidmet. Wählen Sie einfach einen Operator aus und schreiben Sie beliebigen Text hierzu in den Kommentarbereich. Dieser wird dann zusammen mit Ihrer Prozess-Definition gespeichert und kann später nützlich sein, um einzelne Schritte im Design nachvollziehen zu können. Die Tatsache, dass ein Kommentar zu einem Operator vorliegt, wird durch ein kleines Text-Icon am unteren Rand

des Operators angezeigt.

2.3.7 Overview View

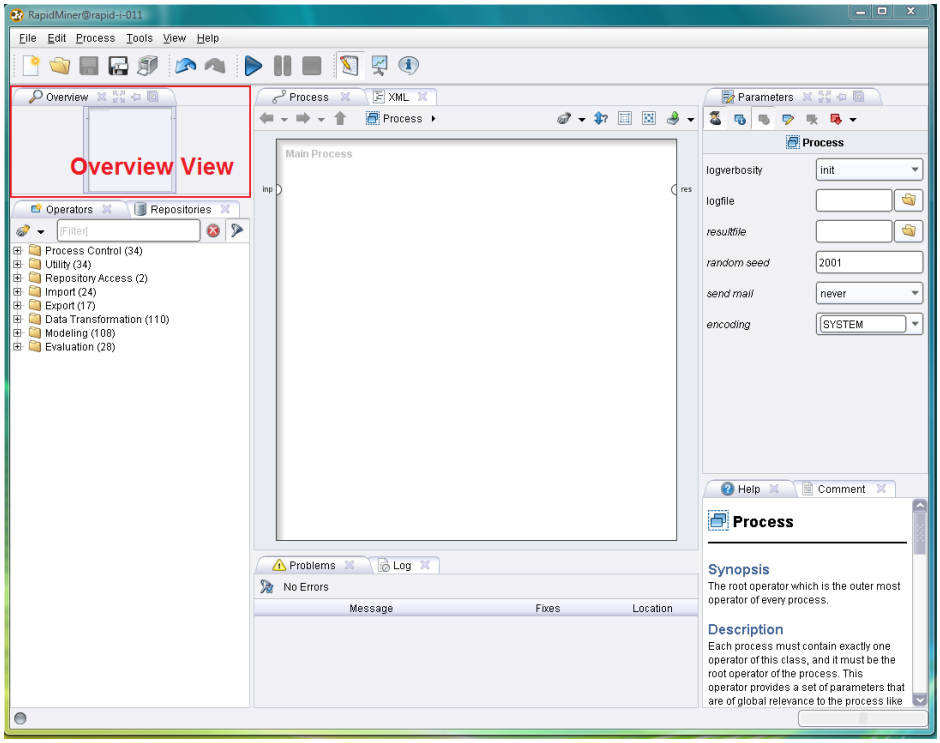


Abbildung 2.28: Behalten Sie den Überblick mittels des Overview Views.

Speziell bei umfangreichen Prozessen wird der weiße Arbeitsbereich nicht mehr ausreichen und man wird diesen entweder mittels des Kontextmenüs des Process Views, mittels der Tastenkombinationen aus STRG und Pfeil links, rechts, oben und unten oder einfach mittels Draggen eines Operators an den Rand vergrößern. In diesem Fall wird jedoch nicht mehr der gesamte Arbeitsbereich gleichzeitig sichtbar sein und eine Navigation innerhalb des Prozesses wird erschwert. Um die Übersicht zu erhöhen und gleichzeitig eine komfortable Möglichkeit zur Navigation zu liefern, bietet RapidMiner den Overview View, welcher den gesamten Arbeitsbereich zeigt und mittels eines kleinen Kastens den aktuell angezeigten Ausschnitt markiert:

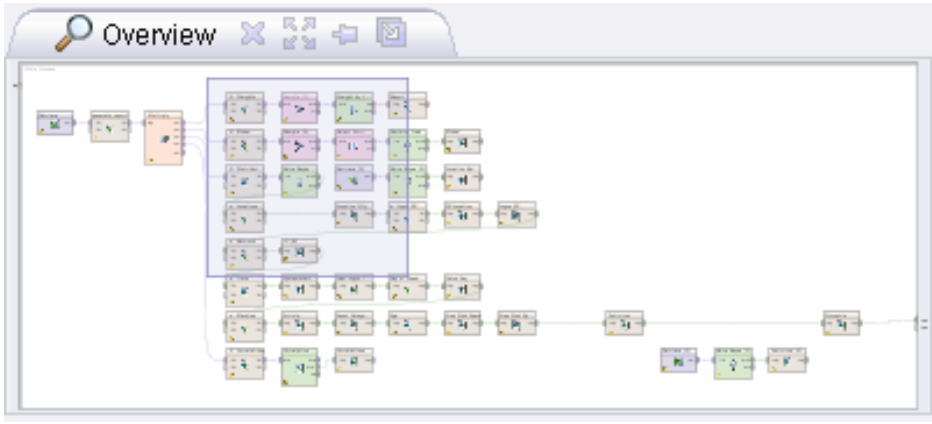


Abbildung 2.29: Der Overview View zeigt den gesamten Prozess und markiert den sichtbaren Ausschnitt.

Sie werden sehen, dass sich der Ausschnitt beim Scrollen innerhalb des Process View verschiebt – und nun mittels der Scrollbar oder einfach durch Draggen eines Operators an den Rand des Ausschnitts. Gleichzeitig können Sie aber auch einfach den markierten Bereich in diesem Overview an die gewünschte Stelle ziehen und der Process View passt sich automatisch an.

2.3.8 Problems und Log View

Abbildung 2.30 zeigt die Problems und Log View von RapidMiner.

Problems View

Ein weiteres ausgesprochen zentrales Element und eine wertvolle Hilfe während des Designs Ihrer Analyseprozesse ist der Problems View. In diesem werden alle Warnungen und Fehlermeldungen übersichtlich in einer Tabelle angezeigt (Abbildung 2.31).

In der ersten Spalte mit dem Namen „Message“ finden Sie eine kurze Zusammenfassung des Problems. In diesem Fall ist das Data Mining Verfahren „Gaussian Process“ nicht in der Lage, polynominale – also mehrwertige kategorielle – Attribute zu behandeln. Die letzte Spalte namens „Location“ gibt Ihnen die Stelle an, an der das Problem auftritt in Form des Operatornamens und des Namens

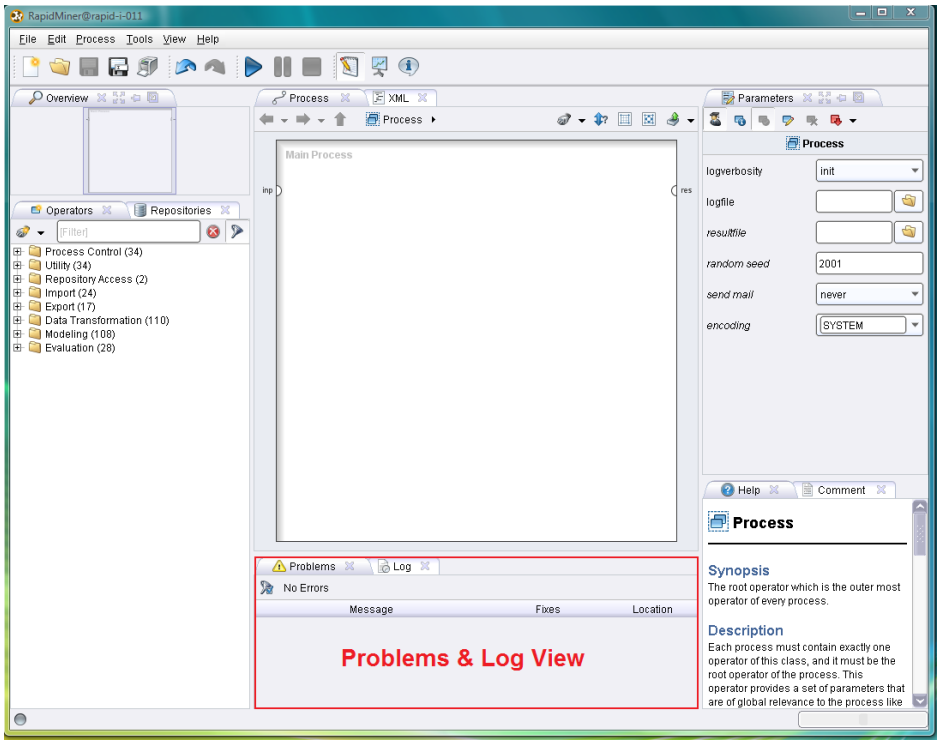


Abbildung 2.30: Die Tabelle im Problems View zeigt alle (potentiellen) Probleme im Design übersichtlich an und gibt in zahlreichen Fällen auch gleich Hinweise zur Lösung (Quick Fixes). Weitere Informationen finden Sie im Log View.

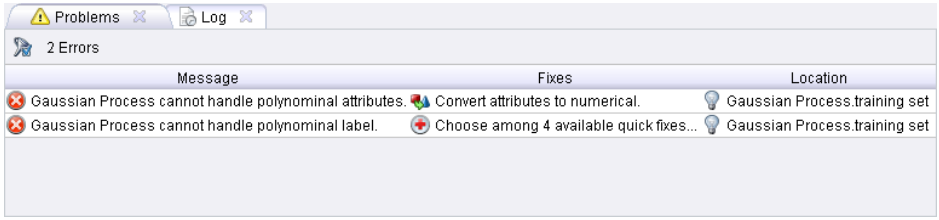


Abbildung 2.31: Darstellung aller aktuellen Probleme.

des betreffenden Input-Ports. Beachten Sie bitte auch das rechts in der Toolbar des Problems View. Hiermit können Sie einen Filter aktivieren, so dass nur noch die Probleme des aktuell ausgewählten Operators angezeigt werden. Dies ist bei größeren Prozesses mit mehreren Fehlerquellen ungemein praktisch.

Eine wesentliche Neuerung von RapidMiner 5 ist jedoch die Möglichkeit, auch Lösungen für solche Probleme vorzuschlagen und auch direkt auszuführen. Diese Lösungswege werden **Quick Fixes** genannt. Die zweite Spalte gibt eine Übersicht über solche mögliche Lösungen, entweder direkt als Text falls es nur eine Lösungsmöglichkeit gibt oder als Angabe, wie viele verschiedene Möglichkeiten existieren, um das Problem zu lösen. In dem Beispiel oben gibt es zwei verschiedene Möglichkeiten, das zweite Problem zu behandeln. Aber warum heißt dieser Lösungsvorschlag „Quick Fix“? Probieren Sie doch einfach in einem solchen Fall mal einen Doppelklick auf das betreffende Quick-Fix-Feld in der Tabelle. Im ersten Fall würde der Lösungsvorschlag direkt ausgeführt und ein betreffender Operator automatisch so konfiguriert und eingefügt, dass die notwendige Vorverarbeitung durchgeführt wird.

Im zweiten Fall mit mehreren Lösungsmöglichkeiten würde ein Dialog erscheinen, der Sie auffordert, den gewünschten Lösungsweg auszuwählen. Nach Auswahl einer der Möglichkeiten würde auch in diesem Fall einer oder mehrere notwendige Operatoren konfiguriert und so eingefügt, dass das Problem nicht länger auftritt. Auf diese Weise können Sie Probleme bereits sehr früh und ausgesprochen komfortabel bereits während des Design-Prozesses erkennen und in wenigen Klicks beheben.

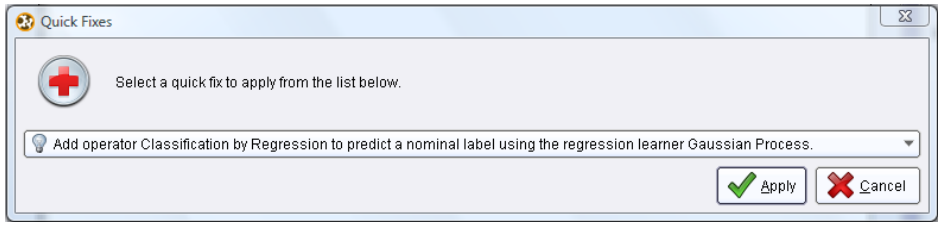

Quick Fixes
Select a quick fix to apply from the list below.
Add operator Classification by Regression to predict a nominal label using the regression learner Gaussian Process.
Apply
Cancel

Abbildung 2.32: Auswahldialog im Falle mehrerer möglicher Quick Fixes.

Hinweis: Die Bestimmung potentieller Probleme wie auch die Generierung von Quick Fixes gehören zu den Funktionen von RapidMiner, die von einer korrekten Bereitstellung von Metadaten abhängig sind. Wir empfehlen Ihnen dringend die Verwendung des Repositories, da andernfalls, beispielsweise bei unmittelbaren Lesen von Daten aus Dateien oder Datenbanken, die Metadaten nicht zur Verfügung stehen und so diese Unterstützungen nicht angeboten werden.

Log View

Während des Designs, aber insbesondere auch während der Ausführung von Prozessen, werden zahlreiche Nachrichten mitgeschrieben und können vor allem im Falle eines Fehlers Aufschluss darüber geben, wie der Fehler durch ein geändertes Prozess-Design behoben werden kann.

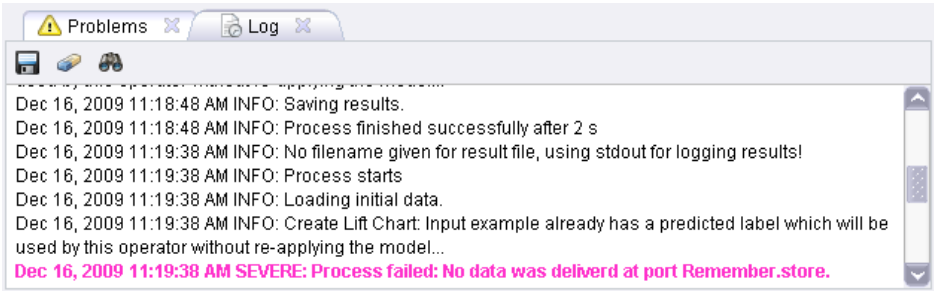

Problems
Log
Dec 16, 2009 11:18:48 AM INFO: Saving results.
Dec 16, 2009 11:18:48 AM INFO: Process finished successfully after 2 s
Dec 16, 2009 11:19:38 AM INFO: No filename given for result file, using stdout for logging results!
Dec 16, 2009 11:19:38 AM INFO: Process starts
Dec 16, 2009 11:19:38 AM INFO: Loading initial data.
Dec 16, 2009 11:19:38 AM INFO: Create Lift Chart: Input example already has a predicted label which will be used by this operator without re-applying the model...
Dec 16, 2009 11:19:38 AM SEVERE: Process failed: No data was deliverd at port Remember.store.

Abbildung 2.33: Weitere Informationen insbesondere zur Prozessausführung und im Fehlerfall finden sich im Log View.

Sie können wie gewohnt den Text innerhalb des Log Views kopieren und in anderen Anwendungen weiter verarbeiten. Sie können mittels der Aktionen in der Toolbar den Text auch in einer Datei speichern, den vollständigen Inhalt löschen oder den Text durchsuchen.

3 Ausführung von Analyseprozessen mit RapidMiner

Wir haben im letzten Kapitel die grundsätzlichen Elemente der graphischen Benutzeroberfläche von RapidMiner wie Perspektiven und Views kennengelernt und die wichtigsten Aspekte der Design-Perspektive von RapidMiner diskutiert. Nun möchten wir die neuen Möglichkeiten dazu nutzen, einen ersten einfachen Analyseprozess zu definieren und auszuführen. Sie werden gleich feststellen, dass es eine äußerst praktische Angelegenheit ist, dass Sie bei RapidMiner den Prozess eben *nicht* für jede Änderung erneut ausführen müssen, um den Effekt der Änderung zu bestimmen. Doch dazu später mehr.

3.1 Erstellen eines neuen Prozesses

Ob Sie nun die Aktion „New“ aus der Welcome-Perspektive wählen, das „New“ Icon ganz links in der Haupt-Toolbar von RapidMiner oder den zugehörigen Eintrag im „File“-Menü: In jedem Fall wird ein neuer Analyseprozess erzeugt, den Sie im Folgenden bearbeiten können. Bevor es jedoch so weit ist, erscheint der „Repository Browser“ (Abbildung 3.1) und fordert Sie auf, einen Speicherort für Ihren neuen Prozess anzugeben.

Wählen Sie einfach ein Repository aus und einen Ort, d.h. ein Verzeichnis, in dem Sie den neuen Prozess speichern möchten. Neue Verzeichnisse können über das Kontextmenü von Repository Einträgen oder auch des Repositorys selbst angelegt werden. Nachdem Sie den Ort gewählt haben, geben Sie Ihrem Prozess

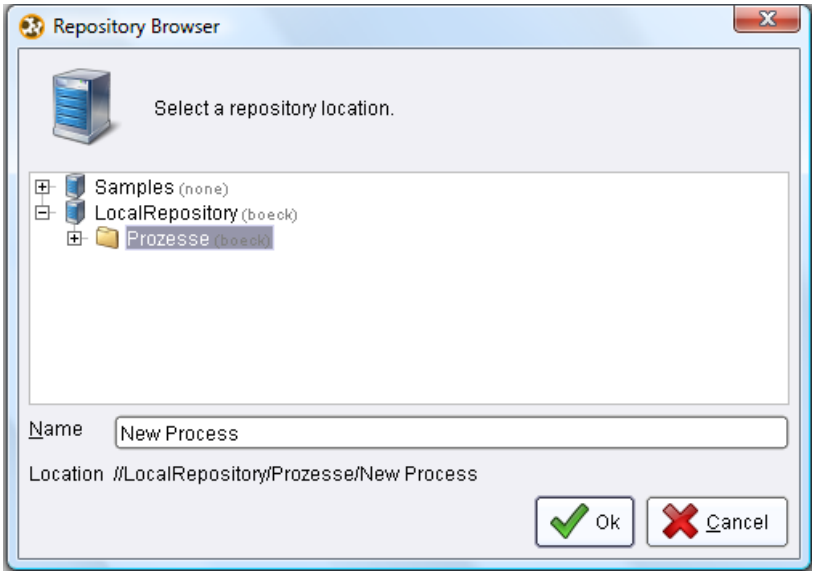


Abbildung 3.1: Der Repository Browser dient zur Auswahl von Repository Einträgen oder Speicherorten analog zu den von Betriebssystemen bekannten Dateidialogen.

noch einen Namen und bestätigen Sie Ihre Wahl durch „Ok“.

Tipp: Sie können auch einen neuen Prozess anlegen, ohne einen Eintrag im Repository zu generieren, indem Sie den Repository Browser mit „Cancel“ schließen. Dies empfiehlt sich jedoch nicht, da die übrigen Repository Einträge wie solche für Daten relativ zu dem Prozess definiert werden. Dies ermöglicht sowohl eine leichtere Ausführung des Prozesses auf Servern im Netzwerk als auch die Weitergabe an andere Analysten beziehungsweise die Kopie für andere Rechner. Wir empfehlen daher, immer auch einen Repository Eintrag für neue Prozesse anzulegen.

Im Prinzip sind Sie völlig frei in der Strukturierung Ihres Repositories. Im Kontextmenü der Einträge im Repository Browser und auch im Repository View finden Sie alle benötigten Einträge zur Verwaltung Ihrer Daten und Prozesse, wie sie in Abbildung 3.2 sehen können.

Diese Aktionen sind im Einzelnen:

1. *Store Process here:* speichert den aktuellen Prozess an den angegebenen

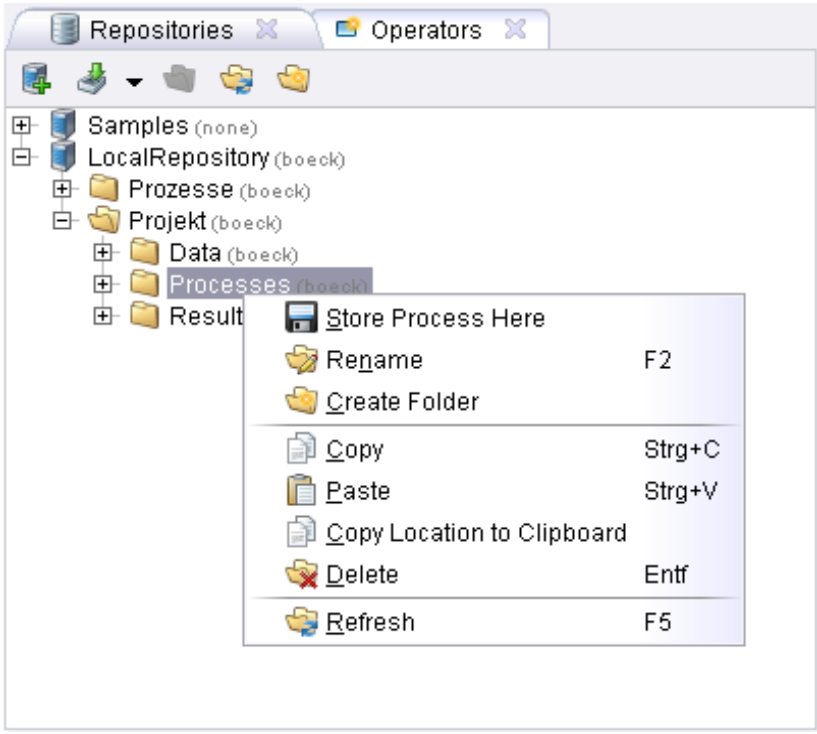


Abbildung 3.2: Das Kontextmenü der Repository-Einträge sowohl im Repository Browser als auch im Repository View bietet alle notwendigen Optionen zur Verwaltung.

Ort,

2. *Rename:* Benennt den Eintrag oder das Verzeichnis um,

3. *Create Folder:* Legt ein neues Verzeichnis an dieser Stelle an,

4. *Delete:* Löscht den gewählten Repository-Eintrag oder Verzeichnis,

5. *Copy:* Kopiert den gewählten Eintrag zum späteren Einfügen an anderen Stellen,

6. *Paste:* Kopiert einen zuvor kopierten Eintrag an diese Stelle,

7. *Copy Location to Clipboard:* Kopiert einen eindeutigen Bezeichner für diesen Eintrag in die Ablage, so dass Sie diese als Parameter für Operatoren, in

Web Interfaces o.ä. nutzen können,

8. *Refresh:* Aktualisiert die Anzeige.

Es empfiehlt sich, für einzelne Analyseprojekte neue Verzeichnisse im Repository anzulegen und diese entsprechend zu benennen. Eine weitere Strukturierung innerhalb der Projekte kann nie schaden, beispielsweise in weitere Unterverzeichnisse für projektspezifische Daten, verschiedene Phasen der Datentransformation und –analyse oder für Ergebnisse. Ein Repository könnte also beispielsweise die folgende Struktur aufweisen:

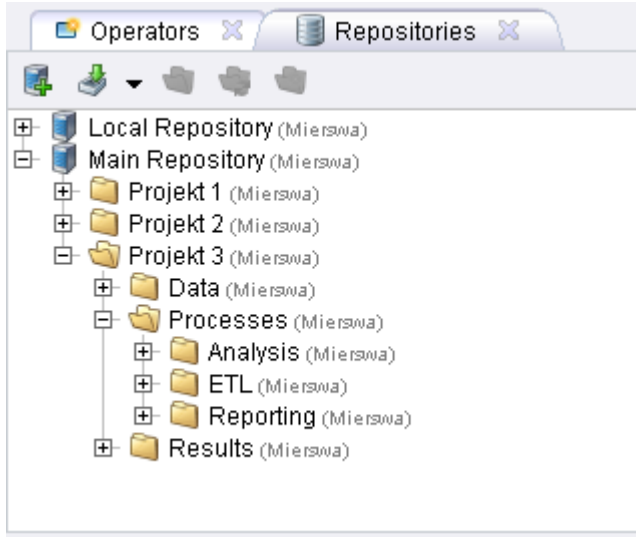


Abbildung 3.3: Ein Repository mit einer Strukturierung in Projekte und dort jeweils nach Daten, Prozessen und Ergebnissen.

3.2 Der erste Analyseprozess

Nachdem Sie den Ort und den Namen des Prozesses definiert haben, wechselt RapidMiner automatisch in die Design-Perspektive und Sie können mit dem Prozess-Design starten. In späteren Kapiteln werden wir uns ausführlich damit beschäftigen, wie Sie Daten in RapidMiner einladen und in Ihrem Repository speichern können. In diesem Abschnitt kommt es uns jedoch eher auf die prinzipielle Ausführung von Prozessen und wir werden daher auf die Analyse echter

Daten noch für einen kurzen Augenblick verzichten.

Sofern Sie die Auswahl und Positionen der einzelnen Views für die Design-Perspektive nicht geändert haben, sollten Sie also in etwa den folgenden Bildschirm vor sich haben:

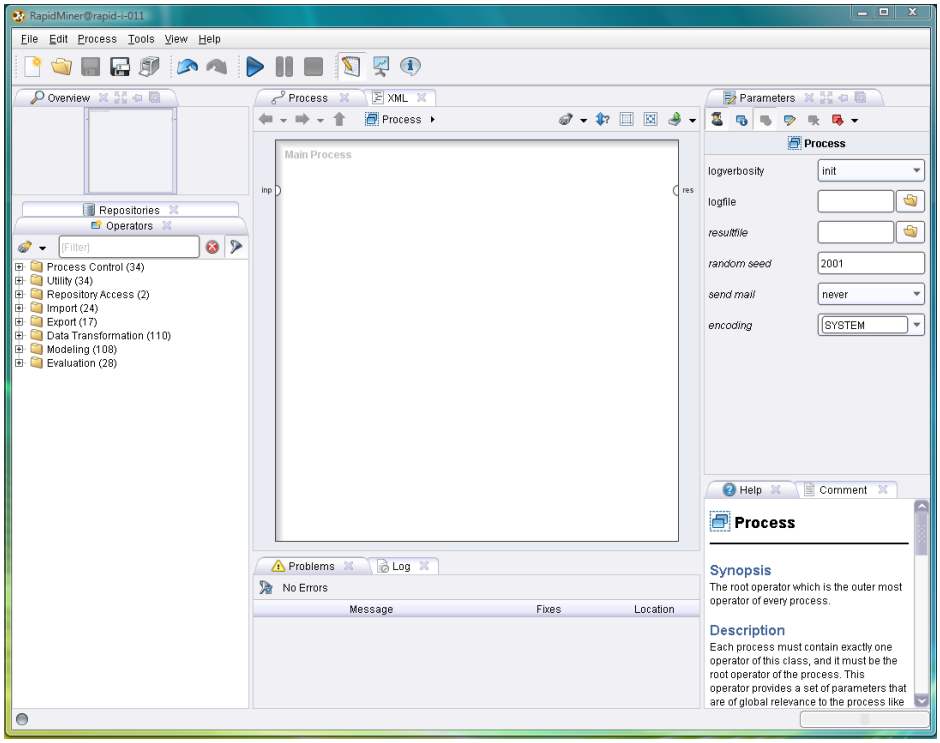


Abbildung 3.4: Die voreingestellte Design-Perspektive unmittelbar nach dem Anlegen eines neuen Prozesses.

Wir beginnen nun unseren neuen Prozess zunächst mit dem Generieren von Daten, auf denen wir arbeiten können. Wie gesagt: In späteren Kapiteln werden wir sehen, wie wir Daten aus dem Repository verwenden können oder mittels Operatoren auch direkt aus anderen Datenquellen wie Datenbanken oder Dateien einlesen. Aber für den Augenblick werden wir darauf verzichten und einen kleinen synthetischen Datensatz generieren.

Klappen Sie nun bitte im Operators View die Gruppe „Utility“ aus und dann „Data Generation“. Die Zahlen in Klammern bei den einzelnen Gruppen geben übrigens die Anzahl an Operatoren für diese Gruppe an. Sie sollten nun mehre-

re Operatoren sehen, die zum Generieren eines künstlichen Datensatzes genutzt werden können. Darunter befindet sich auch der Operator „Generate Sales Data“. Ziehen Sie nun mit gedrückter Maustaste diesen Operator auf die weiße Fläche und lassen Sie dort los. Der Operator wird eingefügt und je nach Einstellung der automatischen Verbindung im Operators View auch direkt verbunden. Falls dies nicht geschieht, können Sie den Output-Port des neuen Operators nun manuell mit dem ersten Ergebnis-Port des gesamten Prozesses am rechten Rand der weißen Arbeitsfläche verbinden. Alternativ wäre es natürlich auch möglich gewesen, den Operator mittels des New Operator Dialogs einzufügen wie im vorigen Kapitel beschrieben. Wie auch immer Sie vorgegangen sind, das Ergebnis müsste nun in etwa so aussehen:

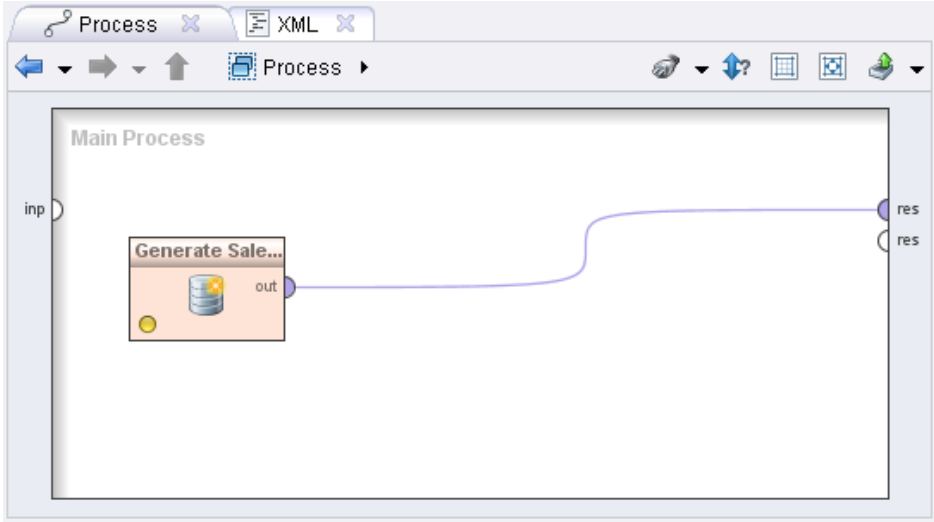


Abbildung 3.5: Ein erster und zugegeben sehr einfacher Prozess, der einige Daten generiert und das Ergebnis in der Result-Perspektive anzeigt.

Wie Sie sicher bemerkt haben, ist der vollständige Name „Generate Sales Data“ dieses Operators zu lang, er wird nach den ersten Buchstaben abgeschnitten. Bewegen Sie die Maus doch einmal auf diesen Operator und verweilen Sie dort für einige Augenblicke. In einer kleinen Animation wird der Name nun vollständig dargestellt. Sie könnten den Operator aber natürlich auch umbenennen und ihm einen kürzeren Namen geben, allerdings würden Sie dann die schicke Animation verpassen:

Wie Sie sehen können, ist die Statusanzeige des Operators unten links gelb.

Abbildung 3.6: Lange Namen werden angezeigt, wenn der Mauszeiger länger auf einem Operator ruhig verweilt.

Dies bedeutet, dass der Operator keine Fehler produziert hat, aber bisher auch noch nicht erfolgreich ausgeführt wurde. Sie haben den Operator also bislang nur vollständig konfiguriert, direkt ausgeführt wurde er deswegen jedoch noch lange nicht. Das könnten Sie leicht daran erkennen, dass die Statusanzeige dann auf Grün wechselt. Sie haben gar nicht bemerkt, dass Sie den Operator bereits konfiguriert haben? Die Konfiguration war in diesem konkreten Fall ja auch denkbar einfach: Es war nämlich gar nicht notwendig, irgendeinen Parameter des Operators einzustellen. Eine rote Statusanzeige und Einträge im Problems View hätten Sie auf solch einen Konfigurationsbedarf hingewiesen.

3.2.1 Transformation der Metadaten

Wir behandeln nun einen der faszinierendsten Aspekte von RapidMiner, nämlich die Fähigkeit, die Ausgabe eines Operators oder eines Prozesses bereits im Vorfeld zu berechnen und dies sogar während der Design-Zeit, also ohne die tatsächlichen Daten laden zu müssen oder den Prozess gar durch zu führen. Dies wird ermöglicht durch die sogenannte *Metadaten-Transformation* von RapidMiner.

Jeder Operator definiert natürlich, auf welche Art und Weise die entgegengenommenen Eingabedaten transformiert werden. Dies ist ja schließlich seine Aufgabe. Das Besondere an RapidMiner jedoch ist, dass dies nicht nur für tatsächliche Daten passieren kann sondern auch für die Metadaten über diese Daten. Diese sind typischerweise deutlich weniger umfangreich als die Daten selbst und geben dem Analysten eine hervorragende Abschätzung darüber, welche Eigenschaften ein bestimmter Datensatz hat. Die Metadaten in RapidMiner entsprechen im Wesentlichen den Konzeptbeschreibungen, die wir bereits früher diskutiert haben. Sie enthalten die Attributnamen der Beispielmenge genauso wie die Wertetypen und die Rollen der Attribute und sogar einige grundlegende Statistiken.

So weit zur Theorie, aber wie sehen die Metadaten in der Praxis, also RapidMiner, aus? In RapidMiner werden die Metadaten an den Ports bereit gestellt. Fahren Sie mit dem Mauszeiger doch einfach mal über den Output-Port des soeben eingefügten Operators und beobachten Sie, was passiert:

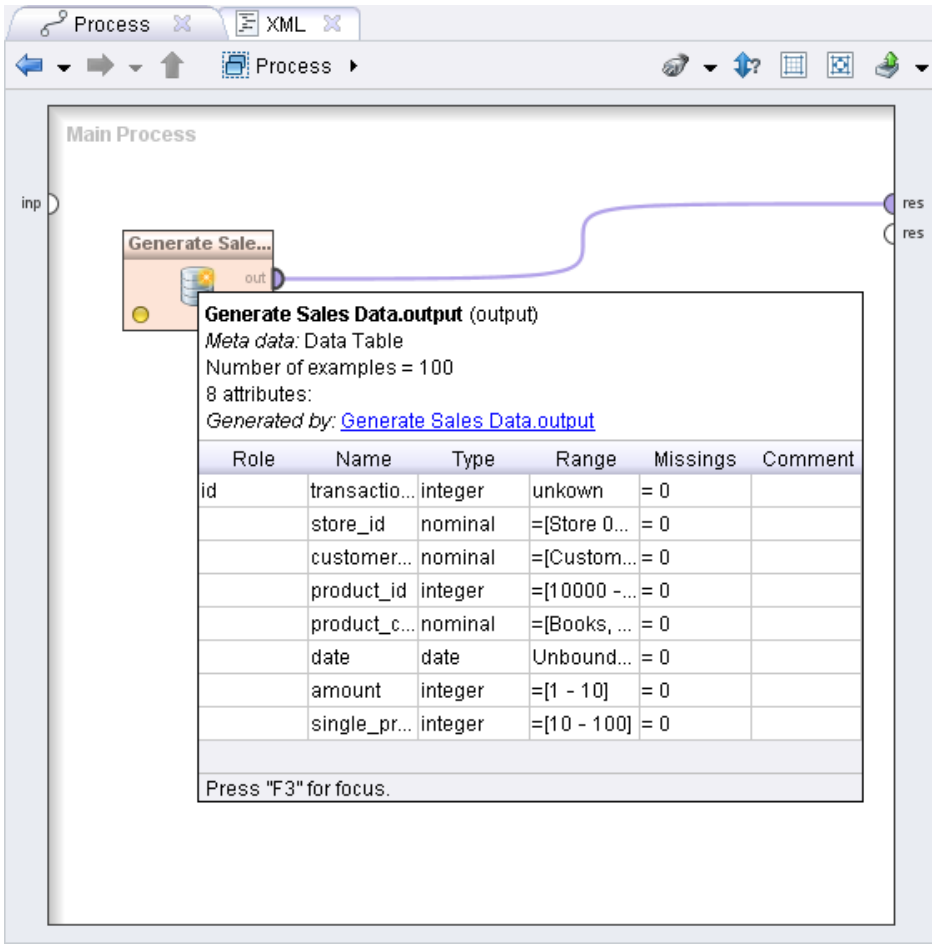


Abbildung 3.7: Die Metadaten des Output-Ports des Operators „Generate Sales Data“.

Es erscheint ein Tooltip, der die erwartete Ausgabe des Ports beschreibt. Zunächst der Name des Operators und des Ports gefolgt von der Art der Metadaten. In diesem Fall handelt es sich um die Metadaten einer Beispielmenge. Die Zahl der Beispiele kann genauso entnommen werden (100) wie die Zahl der Attribute (8). Als nächstes folgt noch eine Beschreibung des Weges, den das Objekt bei einer

Ausführung durch den Prozess absolviert haben würde. In diesem Fall hat der Weg nur eine einzige Station, nämlich den Port des generierenden Operators. Der wichtigste Teil der Metadaten – zumindest für Beispielmenge – ist jedoch die Tabelle, welche die Metadaten der einzelnen Attribute beschreibt. Die einzelnen Spalten sind:

1. *Role:* Die Rolle des Attributs, ohne Angabe handelt es sich um ein reguläres Attribut,
2. *Name:* Der Name des Attributs,
3. *Type:* Der Wertetyp des Attributs,
4. *Range:* Der Wertebereich des Attributs, also Minimum und Maximum bei numerischen Attributen und ein Auszug der möglichen Werte bei nominalen Attributen,
5. *Missings:* Die Zahl der Beispiele, bei denen der Wert dieses Attributs unbekannt ist.

Tipp: Solche komplexeren Tooltips gibt es an mehreren Stellen in RapidMiner, beispielsweise auch für die Operatorbeschreibungen, die als Tooltip im Operators View angezeigt werden. Sie können den Tooltip in aller Ruhe lesen und auch in der Größe anpassen, wenn Sie zuvor die Taste F3 drücken.

Beachten Sie bitte, dass die Metadaten oftmals nur eine Schätzung darstellen können und manchmal eine exakte Angabe nicht möglich ist. Dies äußert sich dadurch, dass Teile der Metadaten unbekannt sind oder nur ungenau angegeben werden können, beispielsweise mit der Angabe „<100 Examples“ für die Zahl der Beispiele. Trotzdem sind die Metadaten eine wertvolle Hilfe sowohl bei den nächsten Designentscheidungen als auch bei der automatischen Erkennung von Problemen sowie den Vorschlägen für deren Lösungen, also den Quick Fixes.

Zurück zu unserem Beispiel. Geschulte Analysten werden auf einen Blick erkennen, dass es sich bei den Daten um sogenannte Transaktionsdaten handeln muss, bei denen jede Transaktion einen Einkauf darstellt. Wir haben für unsere Beispielmenge die folgenden Attribute gegeben:

- *transaction_id:* gibt eine eindeutige ID für die jeweiligen Transaktionen an,
- *store_id:* gibt das Geschäft an, in dem die Transaktion getätigt wurde,

- *customer_id:* gibt den Kunden an, mit dem die Transaktion durchgeführt wurde,
- *product_id:* gibt die ID des gekauften Produkts an,
- *product_category:* gibt die Kategorie des gekauften Produkts an,
- *date:* gibt das Transaktionsdatum an,
- *amount:* gibt die Anzahl der gekauften Objekte an,
- *single_price:* gibt den Preis eines einzelnen Objekts an.

Betrachten wir zunächst die letzten beiden Attribute, so fällt auf, dass zwar die Anzahl und der Einzelpreis der Objekte innerhalb der Transaktion gegeben sind, nicht jedoch der damit verbundene Gesamtumsatz. Als nächstes wollen wir deshalb ein neues Attribut mit Namen „total_price“ generieren, dessen Werte dem Produkt aus Anzahl und Einzelpreis entsprechen. Hierzu verwenden wir einen weiteren Operator namens „Generate Attributes“, der sich in der Gruppe „Data Transformation“ – „Attribute Set Reduction and Transformation“ – „Generation“ befindet. Ziehen Sie den Operator hinter den ersten Operator und verbinden Sie den Output-Port des Datengenerators mit dem Input-Port des neuen Operators sowie dessen Output-Port mit der Ergebnisausgabe des Gesamtprozesses. Es müsste sich etwa das Bild in Abbildung 3.8 ergeben:

Tipp: Statt einen Operator in den Process View zu ziehen und die Ports neu zu verbinden, können Sie den Operator auch auf eine bereits bestehende Verbindung ziehen. Wenn Sie die Position des Mauszeigers genau auf die Verbindung bewegen, wird diese hervorgehoben und der neue Operator direkt in die Verbindung sinnvoll eingefügt.

Auch wenn dieser Prozess nun funktionieren würde, was an den gelben Statusanzeigen und dem leeren Problems View erkannt werden kann, so würde der zweite Operator ohne eine weitere Konfiguration nichts berechnen und das Endergebnis wäre das gleiche wie das nur nach dem ersten Operator. Wir wählen daher den neuen Operator „Generate Attributes“ aus und selektieren ihn auf diese Weise. Die Anzeige im Parameter View ändert sich dementsprechend und die Parameter dieses Operators werden angezeigt. Der wesentliche Parameter hat den Namen „function descriptions“ und wird mit einem Klick, wie in Abbildung 3.9 zu sehen, auf den zugehörigen Knopf konfiguriert:

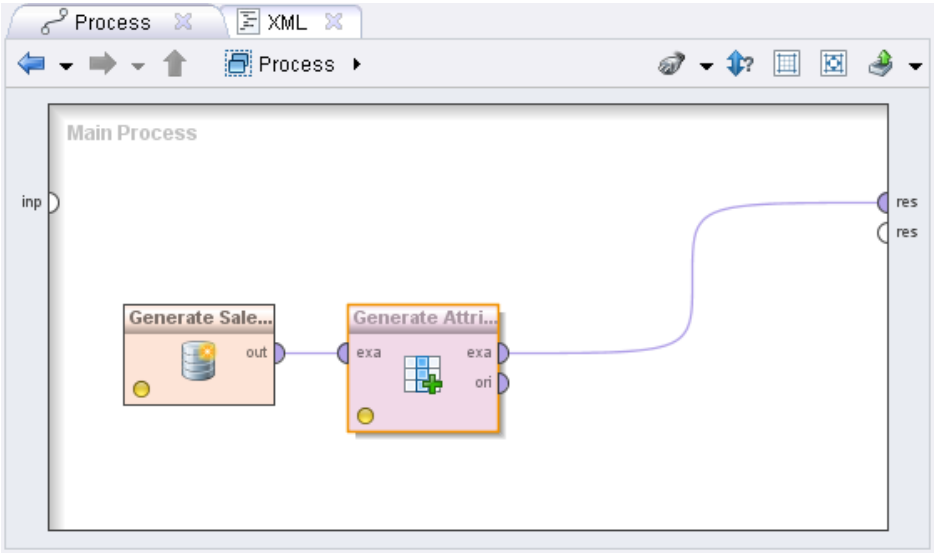


Abbildung 3.8: Die Daten werden zunächst generiert und danach wird ein neues Attribut erzeugt.

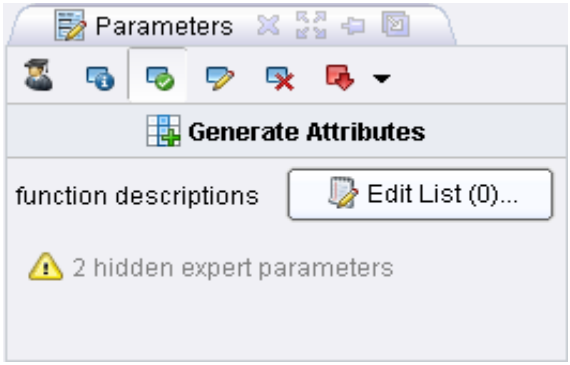


Abbildung 3.9: Die Parameter des Operators „Generate Attributes“.

Nachdem Sie den Knopf mit dem Namen „Edit List (0)“ gedrückt haben, wird sich ein Dialog öffnen, der Ihnen die Gelegenheit gibt, die gewünschte Berechnung in Abbildung 3.10 einzugeben.

Sie können in solchen Listen von Einzelparametern mit den beiden Aktionen „Add Entry“ und „Remove Entry“ weitere Einträge hinzunehmen und ausgewählte Einträge auch wieder löschen. In der Tabellenüberschrift stehen die Namen der

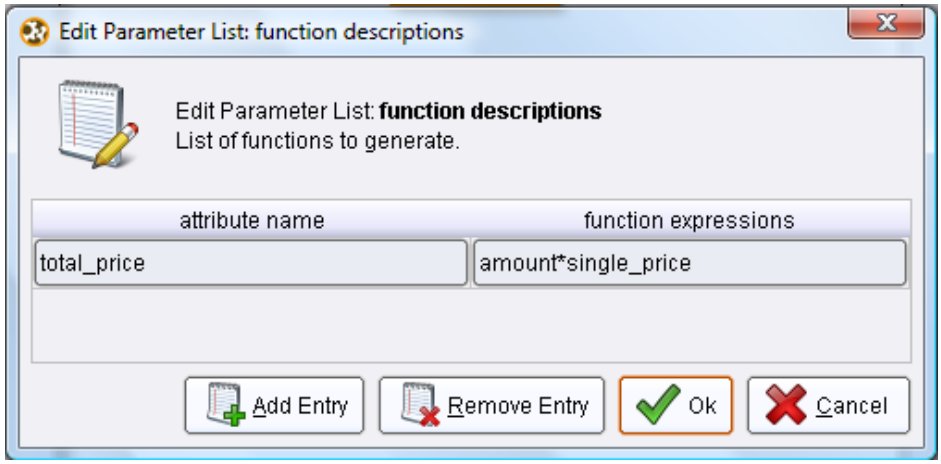


Abbildung 3.10: Berechnung des neuen Attributs „total_price“ als Produkt aus „amount“ und „single_price“.

gewünschten Parameter. Fügen Sie eine Zeile hinzu, geben Sie links den Namen des neuen Attributs ein und rechts die Funktion, die dieses neue Attribut berechnet. In diesem Fall handelt es sich dabei einfach um das Produkt aus zwei anderen Attributen. Bestätigen Sie Ihre Eingabe mit „Ok“ und der Dialog wird sich schließen. Der Knopf mit der Beschriftung „Edit List“ müsste nun in Klammern eine „1“ anzeigen, so dass Sie erkennen können, wie viele Einträge die Parameterliste hat und folglich in diesem Fall auch wie viele neue Attribute generiert werden.

Wir können nun beobachten, wie sich das Hinzufügen des Operator „Generate Attributes“ auf die Metadaten auswirkt. Im Hintergrund hat RapidMiner nämlich bereits die Metadaten transformiert und Sie können sich die neuen Metadaten erneut als Tooltip über den Output-Port des Operators ansehen (Abbildung 3.11).

Es ist in der Zeile „Generatey by“ leicht zu sehen, dass das Objekt nun als letztes dem Operator „Generate Attributes“ entstammt und zuvor dem Operator „Generate Sales Data“. Darüber hinaus hat sich fast nichts geändert, sowohl die Anzahl der Beispiele ist gleich geblieben als auch die acht ursprünglichen Attribute. Es ist jedoch noch ein neuntes Attribut neu hinzugekommen: Unser eben neu definiertes Attribut „total_price“ ist nun ebenfalls in der Tabelle zu finden.

Und noch immer wurde unser Prozess noch nicht ausgeführt, wie Sie leicht an den noch stets gelben Statusanzeigen erkennen können. Sie mögen sich nun viel-

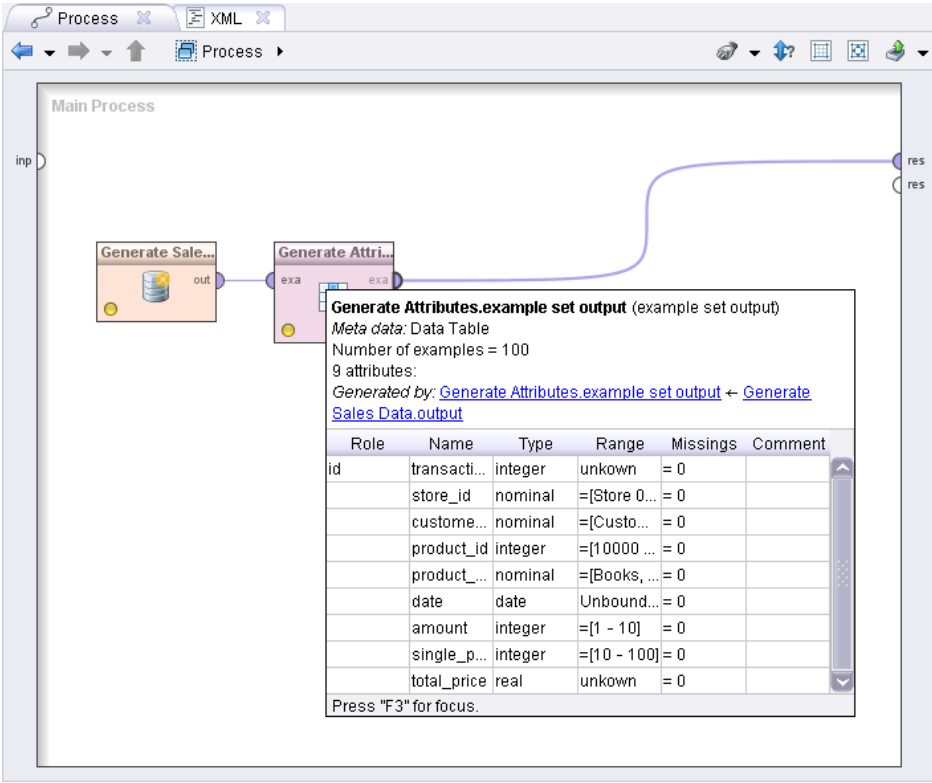


Abbildung 3.11: Die Metadaten enthalten den vollständigen Weg des Objekts und sind bis auf das neu hinzugekommene Attribut „total_price“ unverändert.

leicht fragen: „Na und, ich kenne im Vorfeld das Ergebnis und das ohne Prozessausführung. Was habe ich denn davon?“. Nun, eine ganze Menge. Sie können nun mit einem Blick erkennen, was ein konkreter Operator oder ein (Teil-)Prozess denn mit den Eingabedaten anstellt. Da die Metadaten zudem deutlich kleiner sind als die vollständigen Datensätze, ist diese Überprüfung auch deutlich schneller durchführbar als auf den vollständigen Daten. Sie bekommen auf diese Weise in kürzester Zeit bereits ein Feedback, wenn ein Problem vorliegt, welches vielleicht weitere Datentransformation nötig macht und nicht erst nachdem ein über mehrere Stunden laufender Analyseprozess mit einem Fehler abgebrochen hat. Und zu guter Letzt kann RapidMiner die Informationen aus den Metadaten weiter verarbeiten und Sie beim Design des Prozesses weiter unterstützen, beispielsweise indem bei einer Filterung von Attributen in der graphischen Benutzeroberfläche nur alle noch verfügbaren (und die neu generierten) Attribute angezeigt werden.

Probieren Sie nun zum Beispiel folgendes: Klappen Sie die Gruppe „Data Transformation“ – „Attribute Set Reduction and Transformation“ – „Selection“ auf und ziehen Sie den Operator namens „Select Attributes“ in den Prozess – am besten direkt auf die Verbindung nach dem letzten Operator. Denken Sie daran, dass die Verbindung markiert sein muss bevor Sie den Operator fallen lassen, aber dann wird er direkt korrekt neu verbunden. Sie sollten nun den Prozess wie in Abbildung 3.12 definiert haben.

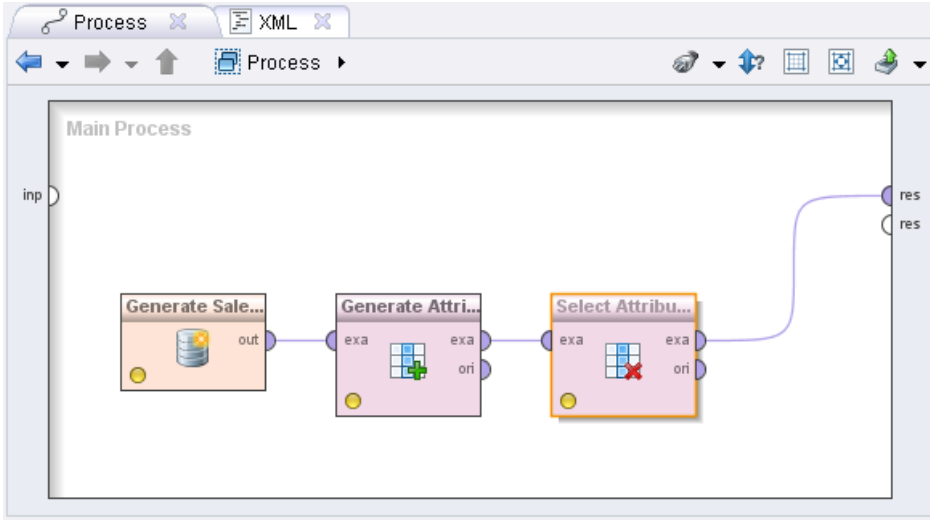


Abbildung 3.12: Generierung der Daten, Generierung eines neuen Attributs, Auswahl einer Teilmenge von Attributen.

Selektieren Sie den neuen Operator und wählen Sie in seinen Parametern für den Parameter „attribute filter type“ die Option „subset“. Beachten Sie bitte, dass nun ein weiterer Parameter namens „attributes“ erschienen ist. Dieser ist fett gedruckt, daher müssen Sie ihn definieren, bevor Sie den Prozess ausführen könnten. Sie erkennen dies auch an der roten Statusanzeige des Operators sowie an dem Eintrag im Problems View. Sie könnten nun den Quick Fix im Problems View per Doppelklick wählen oder auch einfach den Parameter „attributes“ konfigurieren: Erneut per Klick auf einen Knopf, diesmal mit der Beschriftung „Select Attributes...“. Die Parameter sollten wie in Abbildung 3.13 aussehen.

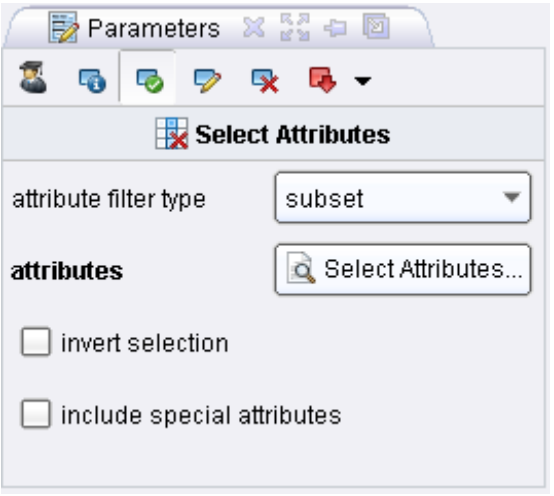


Abbildung 3.13: Der Parameter „attributes“ erscheint nur dann, wenn als Filtertyp „subset“ gewählt wurde.

Drücken Sie nun den Knopf mit der Beschriftung „Select Attributes...“ und wählen Sie in dem erscheinenden Dialog (Abbildung 3.14) aus der Liste entweder per Doppelklick oder per Button mit Pfeil nach rechts in der Mitte die Attribute „product_category“, „store_id“ und „total_price“ aus:

Haben Sie es bemerkt? Das neue und bisher nur im Rahmen der Metadaten-Transformation berechnete Attribut „total_price“ stand Ihnen an dieser Stelle bereits bequem zur Auswahl zur Verfügung – und das ohne, dass Sie den Prozess jemals ausgeführt haben. Wenn Sie die Metadaten am Output-Port erneut überprüfen, so sind nur die drei gewählten Attribute übrig plus die Transaktions-ID, die allerdings auch eine spezielle Rolle – nämlich die der ID – innehat und daher nicht von der Auswahl betroffen war. Da wir diese ID ebenfalls entfernen

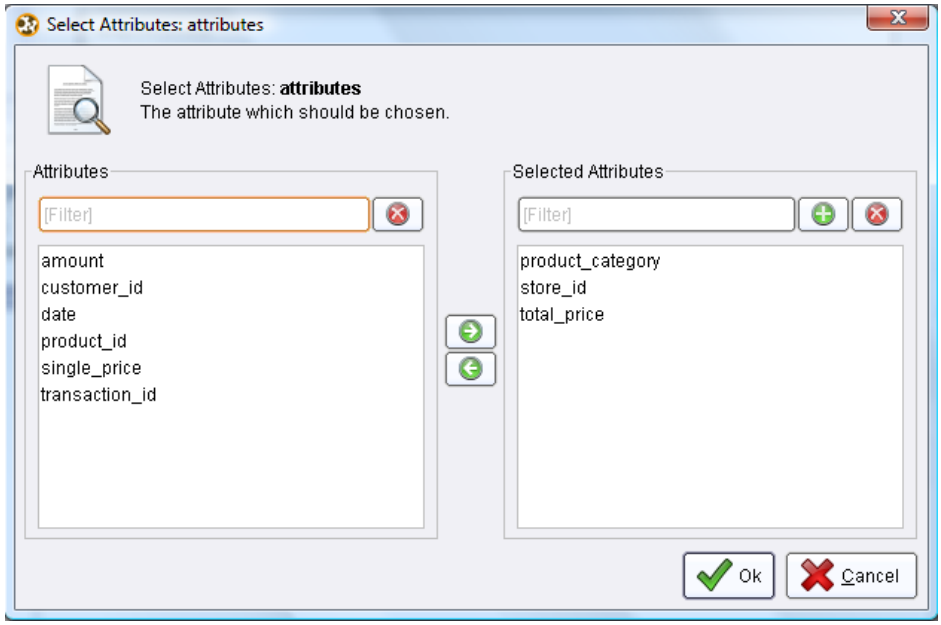


Abbildung 3.14: Mit dem Operator „Select Attributes“ können einzelne Attribute oder Teilmengen ausgewählt oder auch gelöscht werden.

möchten, wählen Sie in den Parametern des Operators „Select Attributes“ die Option „include special attributes“ an und überprüfen Sie die Metadaten erneut: Es sind nun nur noch die drei gewünschten Attribute übrig. Die Auswirkungen dieser und aller anderen Parameter finden Sie in der Beschreibung der Parameter im Help View und auch in der Operator Referenz.

Tipp: Es ist eine Grundregel bei RapidMiner, dass Operatoren aus der Gruppe „Data Transformation“ üblicherweise nur auf regulären Attributen durchgeführt werden, also auf solchen ohne eine spezielle Rolle. Die Operatoren bieten hierzu jedoch eine Option „include special attributes“, so dass sich die Änderungen auch auf diejenigen mit einer besonderen Rolle beziehen.

3.3 Ausführung von Prozessen

Nun sind wir soweit und wir wollen den gerade erstellten Prozess erstmalig ausführen. Die Statusanzeigen aller Operatoren sollten nun gelb sein und es soll-

ten keine Einträge im Problem View existieren. In solch einem Fall sollte unser Prozess, bestehend aus den drei Operatoren zum Generieren der Daten, zur Berechnung des Gesamtumsatzes je Transaktion und zur Filterung von Attributen, problemlos ausführbar sein.

Sie haben zum Starten des Prozesses die folgenden Möglichkeiten:

1. Drücken Sie den großen Play Button in der Toolbar von RapidMiner,
2. Wählen Sie den Menüeintrag „Process“ – „Run“,
3. Drücken Sie F11.

Abbildung 3.15: Der Play-Knopf startet den Prozess, mit dem Pausenknopf können Sie den Prozess zwischenzeitlich anhalten und Stopp bricht den Prozess vollständig ab.

Während ein Prozess läuft, verwandelt sich die Statusanzeige des jeweils gerade ausgeführten Operators in ein kleines grünes Play Icon. Auf diese Weise können Sie erkennen, an welcher Stelle sich der Prozess gerade befindet. Nachdem ein Operator erfolgreich ausgeführt wurde, wechselt die Statusanzeige dann schließlich dauerhaft auf grün – bis Sie bei diesem Operator beispielsweise einen Parameter ändern: Dann zeigt die Statusanzeige erneut eine gelbe Farbe. Das gleiche gilt für alle nachfolgenden Operatoren. So können Sie sehr schnell erkennen, auf welche Operatoren eine Änderung Auswirkungen haben könnte.

Der oben definierte Prozess hat nur eine kurze Laufzeit und daher wird es Ihnen kaum gelingen, den laufenden Prozess zu pausieren oder gar anzuhalten. Prinzipiell jedoch können Sie mit dem Pause-Symbol einen laufenden Prozess kurzzeitig anhalten, beispielsweise um ein Zwischenergebnis anzusehen. Der gerade ausgeführte Operator wird dann noch zu Ende ausgeführt und der Prozess dann angehalten. Sie können einen noch laufenden – aber derzeit angehaltenen – Prozess daran erkennen, dass die Farbe des Play Icons von blau nach grün wechselt. Drücken Sie den Play-Knopf erneut, um den Prozess weiter auszuführen.

Wenn Sie den Prozess nicht nur pausieren, sondern vollständig abbrechen wollen, so können Sie hierzu den Stopp-Knopf betätigen. Genau wie beim Pausieren wird

auch hier der aktuell ausgeführte Operator noch zu Ende durchgeführt und der Prozess direkt im Anschluss vollständig abgebrochen. Bitte beachten Sie, dass Sie direkt nach dem Abbrechen des Prozesses in die Design-Perspektive wechseln können und Änderungen an Prozessen vornehmen – auch wenn der aktuelle Operator im Hintergrund noch zu Ende durchgeführt wird. Sie können sogar weitere Prozesse starten und brauchen nicht auf die vollständige Beendigung des ersten Prozesses zu warten.

Hinweis: Oben wurde darauf hingewiesen, dass der gerade ausgeführte Operator in jedem Fall bei einem Abbruch noch zu Ende ausgeführt wird. Dies ist notwendig, um eine saubere Durchführung von Operatoren zu gewährleisten. Jedoch kann die Fertigstellung eines Operators im Einzelfall noch sehr viel Zeit und auch andere Ressourcen wie Speicherplatz benötigen. Sollten Sie beim Abbruch sehr aufwändiger Operatoren also absehen können, dass dieser beispielsweise noch Stunden laufen wird und die zusätzlichen Ressourcen benötigen, so bleibt Ihnen nur der Neustart der Applikation.

3.3.1 Betrachten von Ergebnissen

Nachdem der Prozess beendet wurde, sollte RapidMiner darauf hingewiesen haben, dass neue Ergebnisse vorliegen und fragen, ob in die Result-Perspektive gewechselt werden soll. War dies bei Ihnen nicht der Fall, so haben Sie wahrscheinlich den Output-Port des letzten Operators nicht mit einem der Ergebnis-Ports des Prozesses am rechten Rand verbunden. Prüfen Sie dies und auch auf andere mögliche Fehler und beachten Sie in diesem Fall die Hinweise im Problems View (Abbildung 3.16).

Sie können sich gerne ein wenig mit den Ergebnissen beschäftigen. Da der obige Prozess noch keine Modellierung durchgeführt hat sondern nur Daten transformiert, besteht das Ergebnis lediglich aus einer Beispielmenge (Example Set). Sie können die Metadaten dieses Datensatzes betrachten, die Tabelle selbst und auch gerne einige der Visualisierungen im Plot View ausprobieren. Im nächsten Kapitel werden wir dann ausführlich die Möglichkeiten der Result-Perspektive behandeln. Wenn Sie wieder in die De-sign-Perspektive zurückkehren wollen, so können Sie dies jederzeit mit den bereits bekannten Mitteln zum Umschalten tun.

Tipp: Nach einiger Zeit werden Sie häufig zwischen Design-Perspektive und Result-Perspektive umschalten wollen. Statt die Icons oder die Menüeinträge zu ver-

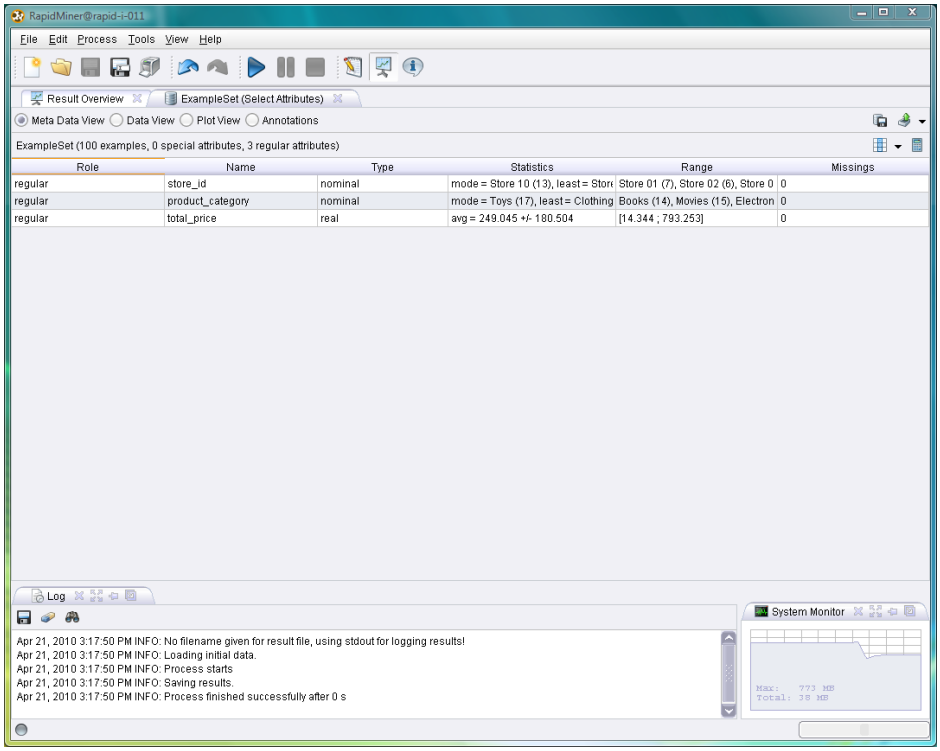

RapidMiner@rapid-i-011
File Edit Process Tools View Help
Result Overview
ExampleSet (Select Attributes)
Meta Data View Data View Plot View Annotations
ExampleSet (100 examples, 0 special attributes, 3 regular attributes)
Role Name Type Statistics Range Missings
regular store_id nominal mode = Store 10 (13), least = Stor Store 01 (7), Store 02 (6), Store 0 0
regular product_category nominal mode = Toys (17), least = Clothing Books (14), Movies (15), Electron 0
regular total_price real avg = 249.045 +/- 180.504 [14.344 ; 793.253] 0
Log
System Monitor
Apr 21, 2010 3:17:50 PM INFO: No filename given for result file, using stdout for logging results!
Apr 21, 2010 3:17:50 PM INFO: Loading initial data.
Apr 21, 2010 3:17:50 PM INFO: Process starts
Apr 21, 2010 3:17:50 PM INFO: Saving results.
Apr 21, 2010 3:17:50 PM INFO: Process finished successfully after 0 s
Max: 773 MB
Total: 38 MB

Abbildung 3.16: Nach erfolgreicher Durchführung eines Prozesses können Sie in der Result-Perspektive die Ergebnisse betrachten.

wenden, können Sie hierzu auch die Tastaturkommandos F8 für einen Wechsel in die Design-Perspektive und F9 für einen Wechsel in die Result-Perspektive verwenden.

3.3.2 Breakpoints

Die Metadaten-Transformation stellt ein sehr mächtiges Werkzeug dar, um das Design von Analyseprozessen zu unterstützen und deutlich komfortabler zu machen. Es entfällt schlicht und ergreifend die Notwendigkeit, den Prozess während des Designs unnötig oft zu Testzwecken durchführen zu müssen. Das erwartete Resultat kann vielmehr anhand der Metadaten bereits abgeschätzt werden. Damit dürfte die Metadatentransformation und –propagierung die Welt der Datenanalyse ein wenig revolutionieren: statt wie bisher jeden Schritt einzeln durchführen zu müssen, um den nächsten Operator konfigurieren zu können, werden die Ergebnisse mehrerer Transformationen nun direkt ganz ohne Ausführung absehbar. Dies ist natürlich insbesondere für die Analyse großer Datenmengen ein gewaltiger Durchbruch.

Trotzdem ergibt sich in einigen Fällen die Notwendigkeit, über die Metadaten hinaus ein konkretes Ergebnis vollständig sehen zu können. Während des laufenden Designs ist es üblicherweise kein Problem, das gewünschte (Zwischen-)Ergebnis an einen Ergebnis-Port des Prozesses zu legen und den Prozess ganz einfach auszuführen. Die gewünschten Ergebnisse werden dann in der Result-Perspektive angezeigt. Aber was können Sie machen, wenn der Prozess bereits fertig designt ist und alle Output-Ports bereits verbunden? Oder sich das Zwischenergebnis tief innerhalb eines verschachtelten Subprozesses befindet? Natürlich gibt es in RapidMiner auch hierfür eine elegante Lösung, die keinerlei Redesign des Prozesses nötig macht. Sie können einfach einen sogenannten Breakpoint einfügen, indem Sie aus dem Kontextmenü eines Operators eine der Optionen „Breakpoint Before“ oder „Breakpoint After“ auswählen, wie in Abbildung 3.17 zu sehen ist.

Wenn ein Breakpoint beispielsweise nach einem Operator eingefügt wurde, so wird die Ausführung des Prozesses an dieser Stelle unterbrochen und die Ergebnisse aller *verbundenen* Output-Ports werden in der Result-Perspektive angezeigt. So können Sie diese Ergebnisse betrachten, ohne dass Sie weitere Änderungen am Prozessdesign vornehmen müssen. Analog zu einem Breakpoint nach einem Operator funktioniert ein Breakpoint vor einem Operator: In diesem Fall wird der

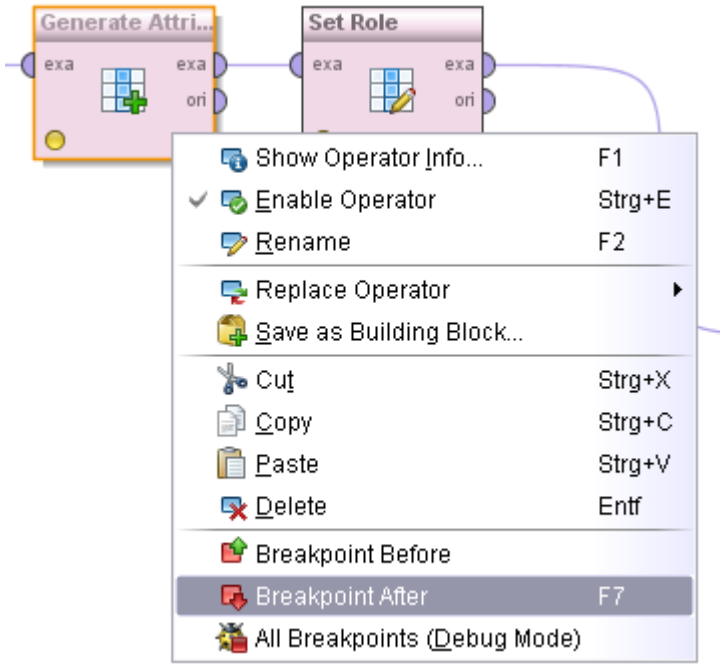


Abbildung 3.17: Mittels Breakpoints können Sie den Prozessablauf anhalten und Zwischenergebnisse inspizieren.

Prozess vor der Ausführung dieses Operators unterbrochen und die Objekte, die an den *verbundenen* Input-Ports dieses Operators anliegen, werden angezeigt. Die Tatsache, dass ein Breakpoint an einem Operator anliegt, wird mittels eines kleinen roten Symbols an der Unterkante des Operators angezeigt (Abbildung 3.18).

Abbildung 3.18: Vor oder nach diesem Operator ist ein Breakpoint definiert.

Tipp: Gerade die Verwendung von „Breakpoint After“ ist relativ häufig, weswegen diese Aktion auch mit einem Tastaturkürzel versehen ist. Mit der Taste

F7 können Sie nach dem derzeitig ausgewählten Operator einen Breakpoint hinzufügen beziehungsweise alle derzeitig vorhandenen Breakpoints entfernen.

Je nachdem, ob Sie RapidMiner entsprechend konfiguriert haben, wechselt RapidMiner automatisch bei einem Breakpoint in die Result-Perspektive und zeigt die Zwischenergebnisse an. Alternativ können Sie einfach selbst in die Result-Perspektive wechseln. Die Tatsache, dass Sie sich zu diesem Zeitpunkt in einem Breakpoint befinden und nicht beispielsweise am Ende des Prozesses, können Sie anhand von zwei Kennzeichen erkennen: Erstens zeigt die Statusanzeige ganz unten links um Hauptfenster von RapidMiner eine rote Ampel, d.h. es läuft zwar ein Prozess, aber er wird derzeit nicht aktiv ausgeführt. Würde derzeit überhaupt kein Prozess laufen, so wäre diese Anzeige einfach grau. Das zweite Kennzeichen für einen Breakpoint ist das nun grüne statt blaue Play-Symbol:

Abbildung 3.19: Das grüne Play-Symbol zeigt an, dass sich der Prozess gerade in einem Breakpoint befindet und durch Pressen wieder weiter ausgeführt werden kann.

Der Prozess kann nun einfach durch Pressen des grünen Play-Symbols wieder aufgenommen werden und zu Ende, oder bis zum nächsten Breakpoint, weiter ausgeführt werden. Natürlich können Sie den Prozess durch Stop wie gewohnt auch vollständig abbrechen.

4 Darstellung von Daten und Ergebnissen

In den vorigen Abschnitten haben wir gesehen, wie die graphische Oberfläche von RapidMiner aufgebaut ist und wie Sie mit ihr Analyseprozesse definieren und ausführen können. Am Ende eines solchen Prozesses können die Ergebnisse des Prozesses dann in der Result-Perspektive angezeigt werden. Wechseln Sie nun mittels eines Klicks in der Toolbar in diese Result-Perspektive. Sie wird im Rahmen dieses Kapitels ausführlich behandelt. Je nachdem, ob Sie bereits darstellbare Ergebnisse erzeugt haben, sollten Sie nun zumindest in den ursprünglichen Einstellungen ungefähr den Bildschirm wie in Abbildung 4.1 vor sich sehen.

Falls nicht, können Sie wie gehabt unter „View“ – „Restore Default Perspective“ diese voreingestellte Perspektive wieder herstellen. Bei der Ergebnisansicht handelt es sich um die zweite zentrale Arbeitsumgebung von RapidMiner neben der bereits besprochenen Design-Perspektive. Der Log-View unten und das Repository rechts oben haben wir bereits zuvor besprochen. In diesem Kapitel werden wir uns daher auf die übrigen Komponenten der Perspektive konzentrieren.

4.1 Systemmonitor

Beim Systemmonitor, den Sie in der voreingestellten Perspektive unten rechts finden, handelt es sich um einen einfachen Speichermonitor, der Ihnen einen Überblick über den gerade verwendeten Speicher gibt. Obwohl RapidMiner bereits durch zahlreiche Maßnahmen, wie beispielsweise der Verzicht auf Datenkopien und stattdessen der Verwendung von Views, versucht, den Speicherbedarf zu reduzieren, so bleibt die Datenanalyse noch stets in vielen Fällen ein Feld mit

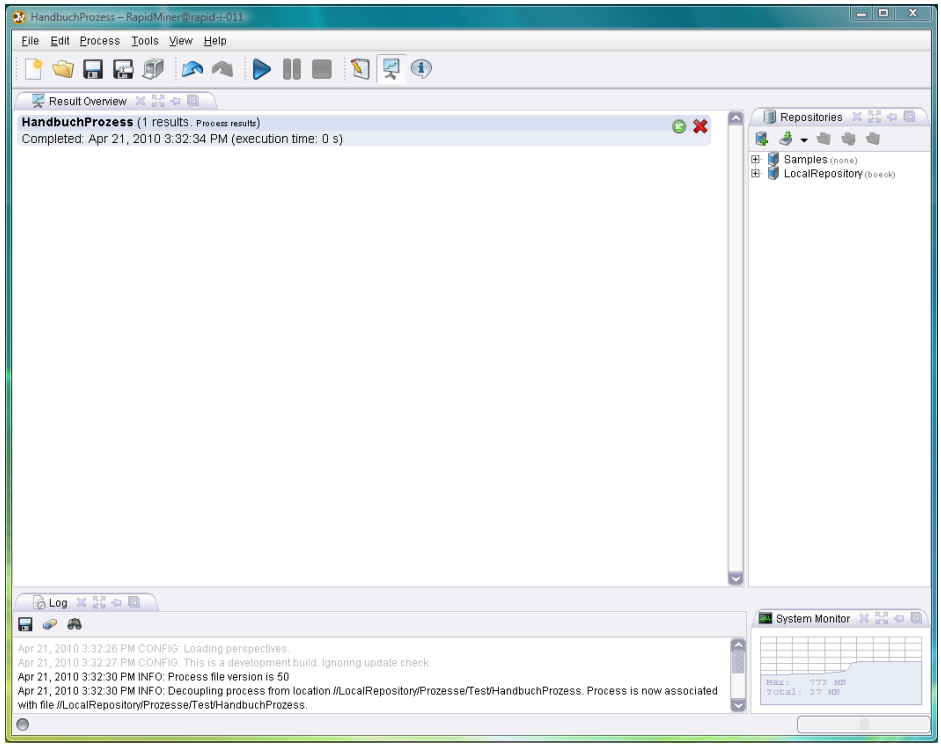

HandbuchProzess – RapidMiner@rapid-i-011
File Edit Process Tools View Help
Result Overview
HandbuchProzess (1 results. Process results)
Completed: Apr 21, 2010 3:32:34 PM (execution time: 0 s)
Repositories
Samples (none)
LocalRepository (boeck)
Log
System Monitor
Apr 21, 2010 3:32:26 PM CONFIG: Loading perspectives.
Apr 21, 2010 3:32:27 PM CONFIG: This is a development build. Ignoring update check.
Apr 21, 2010 3:32:30 PM INFO: Process file version is 50
Apr 21, 2010 3:32:30 PM INFO: Decoupling process from location //LocalRepository/Prozesse/Test/HandbuchProzess. Process is now associated with file //LocalRepository/Prozesse/Test/HandbuchProzess.
Max: 773 MB
Total: 37 MB

Abbildung 4.1: Result-Perspektive von RapidMiner

hohem Speicherbedarf. Der Speichermonitor zeigt Ihnen den maximal in RapidMiner zur Verfügung stehenden Speicher an („Max“) und den höchsten derzeit verwendbaren Speicher („Total„). Letzterer entspricht der oberen Linie des Monitors und kann maximal bis zum absoluten Maximum „Max“ bei Bedarf erhöht werden. Dies geschieht automatisch und nach Möglichkeit nur bei Bedarf. Ist der Speichermonitor vollständig gefüllt, so wird also die bei „Total“ angegebene Menge verwendet. Ist diese genauso hoch wie „Max“, so befindet sich RapidMiner am absoluten Limit und müsste bei noch mehr Speicherbedarf den Prozess abbrechen.

Es ist oftmals möglich, einen solchen Prozess durch geschickte Vorverarbeitung, stapelweiser Bearbeitung, Verwendung von Views oder einem geschicktem Speichermanagement innerhalb von RapidMiner doch noch durchzuführen. Dies ist jedoch ein Feld für Spezialisten und daher nicht Teil dieses Benutzerhandbuchs.

4.2 Anzeigen von Ergebnissen

Wir haben bereits gesehen, dass Objekte, die an die Ergebnis-Ports am rechten Rand eines Prozesses angelegt werden, nach Beendigung des Prozesses automatisch in der Result-Perspektive angezeigt werden. Hierzu dient der große Bereich oben links, in dem auch bereits die Result Overview angezeigt wird, die wir am Schluss dieses Kapitels besprechen werden.

Jedes derzeit geöffnete und angezeigte Ergebnis wird als zusätzliche Registerkarte in diesem Bereich angezeigt:

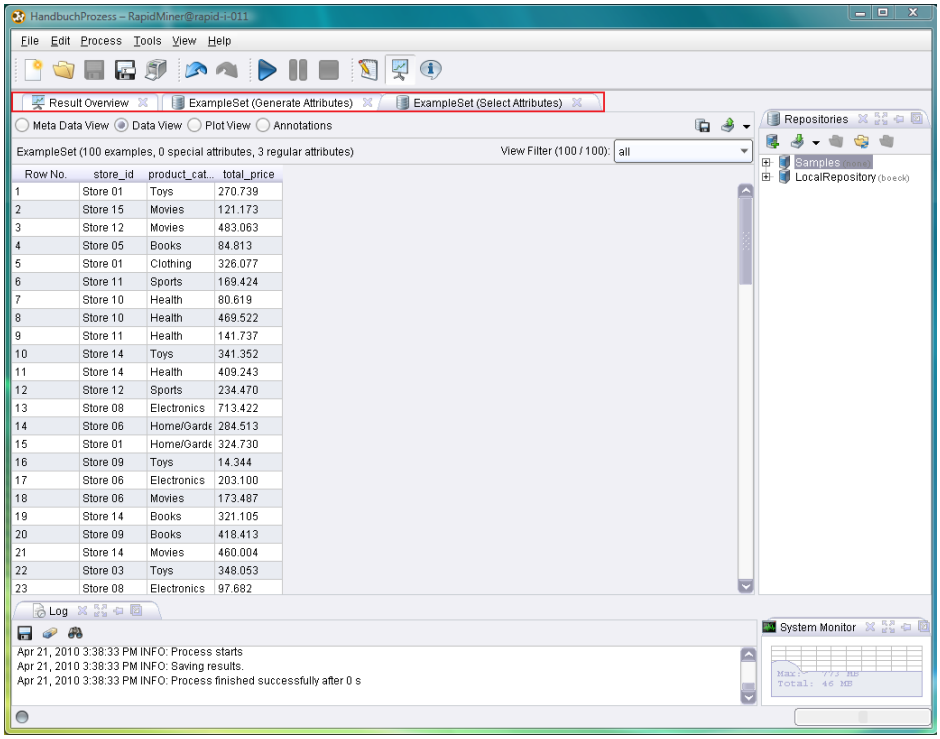

HandbuchProzess – RapidMiner@rapid-i-011
File Edit Process Tools View Help
Result Overview
ExampleSet (Generate Attributes)
ExampleSet (Select Attributes)
Meta Data View
Data View
Plot View
Annotations
ExampleSet (100 examples, 0 special attributes, 3 regular attributes)
View Filter (100 / 100):
all
Row No. store_id product_cat... total_price
1 Store 01 Toys 270.739
2 Store 15 Movies 121.173
3 Store 12 Movies 483.063
4 Store 05 Books 84.813
5 Store 01 Clothing 326.077
6 Store 11 Sports 169.424
7 Store 10 Health 80.619
8 Store 10 Health 469.522
9 Store 11 Health 141.737
10 Store 14 Toys 341.352
11 Store 14 Health 409.243
12 Store 12 Sports 234.470
13 Store 08 Electronics 713.422
14 Store 06 Home/Garde 284.513
15 Store 01 Home/Garde 324.730
16 Store 09 Toys 14.344
17 Store 06 Electronics 203.100
18 Store 06 Movies 173.487
19 Store 14 Books 321.105
20 Store 09 Books 418.413
21 Store 14 Movies 460.004
22 Store 03 Toys 348.053
23 Store 08 Electronics 97.682
Repositories
Samples
LocalRepository (boeck)
Log
Apr 21, 2010 3:38:33 PM INFO: Process starts
Apr 21, 2010 3:38:33 PM INFO: Saving results.
Apr 21, 2010 3:38:33 PM INFO: Process finished successfully after 0 s
System Monitor

Abbildung 4.2: Jedes offene Ergebnis wird als zusätzliche Registerkarte in dem großen Bereich oben links angezeigt.

Bei jedem Ergebnis handelt es sich genau genommen ebenfalls um einen View, den Sie wie gewohnt an beliebige Stellen verschieben können. Auf diese Weise ist es möglich, auch mehrere Ergebnisse gleichzeitig betrachten zu können. Natürlich können Sie auch einzelne Views, d.h. Registerkarten, durch einen Klick auf das

Kreuz in der Karte schließen. Auch die anderen Funktionalitäten von Views wie Maximierung durch Doppelklick etc. stehen Ihnen an dieser Stelle vollständig zur Verfügung.

Sofern Sie die Nachfrage nicht deaktiviert haben, fragt Sie RapidMiner bei Beendigung eines Prozesses, ob die alten Ergebnisse vor Anzeige der neuen Ergebnisse geschlossen werden sollen. Es bleibt letztendlich Ihrem Geschmack überlassen, ob Sie zwecks Vergleichbarkeit alte Ergebnisse prinzipiell offen lassen und manuell schließen wollen. Dank der bereits erwähnten Results Overview scheint diese zusätzliche Arbeit jedoch kaum nötig und so empfehlen wir eher das automatische Schließen der alten Ergebnisse, um die Übersicht zu erhöhen und Verwirrungen auszuschließen.

4.2.1 Quellen für die Anzeige von Ergebnissen

Es gibt mehrere Quellen, aus denen Sie die Anzeige von Ergebnissen speisen können. Wir werden Ihnen im Folgenden alle Möglichkeiten vorstellen:

1. Automatisches Öffnen

Wir haben bereits gesehen, dass die Endresultate eines Prozesses, also solche Objekte, die an die Ergebnis-Ports rechts im Prozess geliefert werden, automatisch angezeigt werden. Gleiches gilt auch für die Ergebnisse an verbundenen Ports im Falle eines Breakpoints. Dies stellt sicher die am häufigsten verwendete und auch empfohlene Variante zur Anzeige von Ergebnisse dar. Sie können einfach alle Ergebnisse an den Ergebnis-Ports des Prozesses sammeln, die Sie am Ende eines Analyseprozesses sehen wollen und alle zusammen werden in den Registerkarten der Result-Perspektive dargestellt.

2. Ergebnisse aus Repositories

Die zweite Möglichkeit zur Anzeige von Ergebnissen ist das Laden von Ergebnissen aus einem Ihrer Repositories. Sie können dies mittels des Kontextmenüs eines Repository-Eintrags oder simpel per Doppelklick auf einen Eintrag bewirken. Dieses Vorgehen ist natürlich nicht nur für die erneute Betrachtung von Ergebnissen empfehlenswert, sondern auch zum Vergleich mit früheren Resultaten.

3. Ergebnisse aus Ports

Eine dritte Möglichkeit, sich Ergebnisse und auch Zwischenergebnisse ansehen zu können, ist die Anzeige von Ergebnissen, welche noch an Ports anliegen. RapidMiner versucht, die Ergebnisse, welche einzelne Operatoren geliefert haben, noch eine zeitlang an den betreffenden Ports zu speichern. Wenn an einem Port noch Ergebnisse anliegen, so können diese über das Kontextmenü des Ports ausgewählt und betrachtet werden:

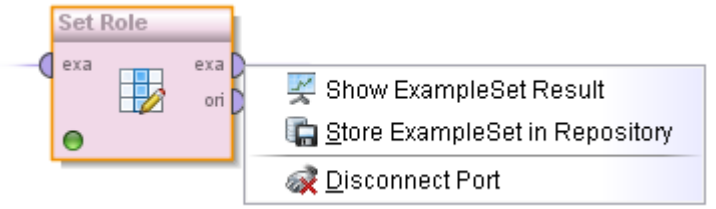


Abbildung 4.3: Anzeige von Ergebnissen, welche noch an Ports anliegen.

Sie kennen diese Vorgehensweise vielleicht von anderen Datenanalysetools: Sie fügen einen Operators hinzu, führen ihn aus und zeigen die Ergebnisse mittels Kontextmenü beziehungsweise mittels spezieller Operatoren hierfür an. Auch wenn diese Vorgehensweise für kleine Datensätze intuitiv und leicht bedienbar schein, so möchten wir dringend von dieser Arbeitsweise abraten, da Sie spätestens bei der Analyse großer Datenmengen zu Problemen führt. In diesem Fall müsste nämlich an jedem Port eine Kopie der Daten vorgehalten werden, um dieses Ergebnis auch später noch zur Verfügung stellen zu können. RapidMiner geht hier einen ganz anderen und langfristig auch erfolgversprechenderen Weg: Die Metadaten werden transformiert und durch den Prozess propagiert und Daten werden nur dort bereitgestellt, wo dieses absolut notwendig ist. Diese Art der RapidMiner-Analyse kombiniert also die Interaktivität, welche durch bekannte Metadaten erlaubt wird mit der einfachen Prozessdefinition für die Analyse auch großer Datenmengen.

Hinweis: RapidMiner besitzt an dieser Stelle ein raffiniertes Speichermanagement. Wie oben bereits erwähnt, werden Ergebnisse noch eine „zeitlang“ an den Ports behalten. Diese Ergebnisse werden gelöscht, sobald der hierfür notwendige Speicher von RapidMinder oder anderen Programmen benötigt wird. Das heißt: Ergebnisse können von den Ports verschwinden und stehen dann auch nicht mehr für eine Visualisierung bereit. Dies ist einer der Gründe für die Effizienz von Ra-

pidMiner und auch aus diesem Grund empfehlen wir die automatische Anzeige über verbundene Ports wie oben beschrieben, da hier die Bereitstellung der Ergebnisse garantiert ist.

4.3 Über Datenkopien und Views

Die Tatsache, dass keine unnötigen Datenkopien angelegt werden, ist manchmal Quelle für Verwirrungen. Dies gilt insbesondere für die oben erwähnte zweite Möglichkeit der Darstellung von Ergebnissen über das Kontextmenü von Ports. Nehmen wir an, Sie haben einen Datensatz und fügen einen Operator für eine Normalisierung hinzu. In seiner Voreinstellung ändert der Normalisierungsoperator die zu Grunde liegenden Daten. Selbst wenn Sie den Datensatz an einem Port betrachten, der im Prozessfluss vor der Normalisierung liegt, aber zeitig nachdem die Normalisierung bereits durchgeführt wurde, so werden sich auch die Daten am Port zuvor bereits geändert haben. Eigentlich sollte dieses Verhalten ausreichend klar sein, es wurde ja wie bereits erwähnt auch keine Kopie der Daten angelegt und der gleiche Datensatz wurde weiter verändert. Und dennoch führt dieses „seltsame“ Verhalten von „unkontrollierten Datenänderungen“ von Zeit zu Verwirrungen.

Sie haben jedoch zwei Möglichkeiten, dieses Verhalten zu beeinflussen:

1. *Verwendung von Views*: Zahlreiche Operatoren für Datentransformationen bieten einen Parameter „create view“, der veranlasst, dass statt einer Änderung der Daten lediglich eine weitere Sicht auf die Daten gelegt wird, die die Daten on-the-fly, also während des Datenzugriffs, ändert. Diese Berechnungen betreffen dann vorherige Ports oder auch Ports in anderen, parallelen Strängen des Prozesses nicht.

2. *Explizite Kopien*: Speziell für kleinere Datensätze kann die Kombination der Operatoren „Multiply“ mit „Materialize Data“ einen Ausweg darstellen. Hiermit definieren Sie als Analyst explizit den Wunsch nach einer Kopie der Daten, indem Sie zunächst die Referenz auf den Datensatz mittels „Multiply“ vervielfältigen und dann beide virtuellen Datensätze explizit mittels „Materialize Data“ als Tabellen neu anlegen.

Kein Analyst wird diesen Aufwand ernsthaft betreiben, lediglich um über die

Ports auf die Ergebnisse zugreifen zu können. Aber auch in parallelen Strängen von Prozessen können solche Querbeziehungen von Zeit zu Zeit auftreten und dann je nach Größe des Datensatzes mittels Views oder auch expliziten Kopien aufgelöst werden.

4.4 Darstellungsformen

Wie auch immer die Ergebnisse in die Result-Perspektive gekommen sind, jedes Ergebnis wird innerhalb einer eigenen Registerkarte angezeigt. Und darüber hinaus, existieren für eine Vielzahl von Ergebnissen noch verschiedene Anzeigemöglichkeiten, die innerhalb von RapidMiner ebenfalls als Views bezeichnet werden:

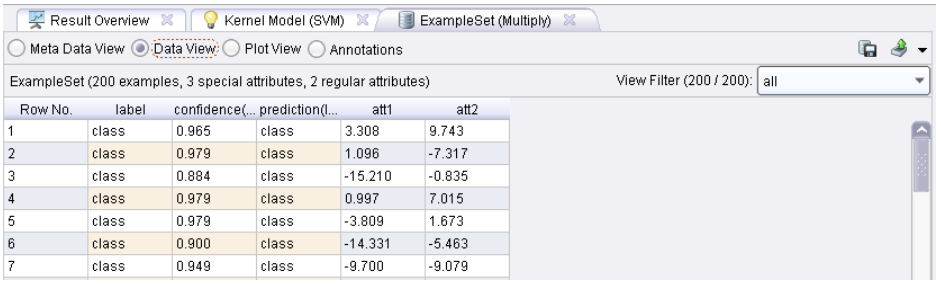

Result Overview
Kernel Model (SVM)
ExampleSet (Multiply)
Meta Data View
Data View
Plot View
Annotations
ExampleSet (200 examples, 3 special attributes, 2 regular attributes)
View Filter (200 / 200):
all
Row No. | label | confidence(... | prediction(l... | att1 | att2
1 | class | 0.965 | class | 3.308 | 9.743
2 | class | 0.979 | class | 1.096 | -7.317
3 | class | 0.884 | class | -15.210 | -0.835
4 | class | 0.979 | class | 0.997 | 7.015
5 | class | 0.979 | class | -3.809 | 1.673
6 | class | 0.900 | class | -14.331 | -5.463
7 | class | 0.949 | class | -9.700 | -9.079

Abbildung 4.4: Für einen Datensatz existieren die Views „Meta Data View“, „Data View“ (derzeit angezeigt) und „Plot View“.

Für Datensätze existieren beispielsweise drei Views, nämlich die Anzeige der Metadaten und Statistiken („Meta Data View“), die Anzeige der Daten selbst („Data View“) sowie die Anzeige von verschiedenen Visualisierungen („Plot View“). Im Beispiel oben sehen Sie die Data View eines Datensatzen in Form einer Tabelle. Neben solchen Tabellen stehen weitere Standard-Darstellungsformen zur Verfügung, die wir im Folgenden erläutern möchten.

Beachten Sie zuvor bitte, dass alle Views sich zwei gemeinsame Schaltflächen oben rechts teilen: das linke Icon dient zum Abspeichern dieses Ergebnisses im Repository und das zweite dient verschiedenen Form des Exports des Ergebnisses, beispielsweise durch Ausdrucken oder Exportieren in eine Grafikdatei.

4.4.1 Text

Die grundlegendste Form der Visualisierung ist die in Form eines Textes. Einige Modelle aber auch zahlreiche andere Ergebnisse können in textueller Form dargestellt werden, typischerweise geschieht dies im Rahmen des sogenannten „Text Views“, den Sie – falls es mehrere Views für dieses Objekt gibt – über die Schaltflächen direkt unterhalb der Registerkarte auswählen können.

In RapidMiner können Sie solche Texte stets mit der Maus markieren und mit STRG + C in die Zwischenablage kopieren. Damit stehen die Ergebnisse dann auch in anderen Applikationen bereit. Längere Texte können Sie mittels eines Klicks auf die Textfläche gefolgt von STRG + A auch vollständig markieren und dann kopieren.

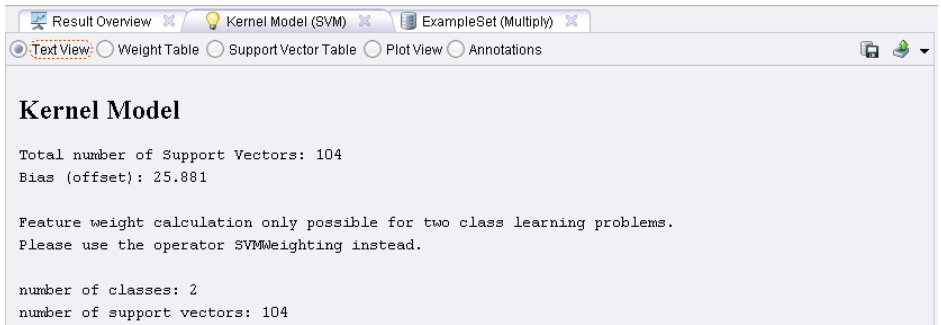


Abbildung 4.5: Einige Modelle wie beispielsweise Regelmengen, werden in textueller Form dargestellt. Aber auch zahlreiche andere Objekte bieten eine Darstellung in Form eines lesbaren Textes.

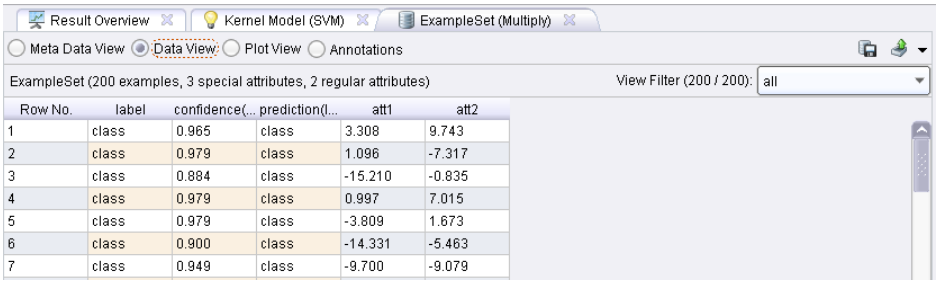
4.4.2 Tabellen

Eine der häufigsten Darstellungsformen von Informationen innerhalb von RapidMiner ist die Form der Tabelle. Dies muss bei einer Softwarelösung, deren vorrangiges Ziel die Analyse von Daten in tabellenartigen Strukturen ist, natürlich auch kaum wundern. Tabellen werden aber nicht nur für die Darstellung von Datensätzen verwendet, sondern auch für die Darstellung von Metadaten, von Gewichten von Einflusseinfaktoren, für die Darstellung von Matrizen wie den Korrelationen zwischen allen Attributen und für vieles andere mehr. Häufig haben diese Ansichten den Begriff „Table“ im Namen, insbesondere wenn Verwechs-

lungen zu befürchten sind. Ansonsten wird schlicht auch über Begriffe wie „Data View“ oder „Meta Data View“ auf solche Tabellen hingewiesen.

Farbschemata

Fast alle Tabellen in RapidMiner nutzen bestimmte Farbkodierungen, die die Übersicht erhöhen. Für Datensätze beispielsweise werden die Zeilen alternierend in unterschiedlichen Farben dargestellt. Attribute mit einer speziellen Rolle erhalten hierbei einen hellgelben Hintergrund und reguläre Attribute einen hellblauen:



Row No.	label	confidence(...	prediction(l...	att1	att2
1	class	0.965	class	3.308	9.743
2	class	0.979	class	1.096	-7.317
3	class	0.884	class	-15.210	-0.835
4	class	0.979	class	0.997	7.015
5	class	0.979	class	-3.809	1.673
6	class	0.900	class	-14.331	-5.463
7	class	0.949	class	-9.700	-9.079

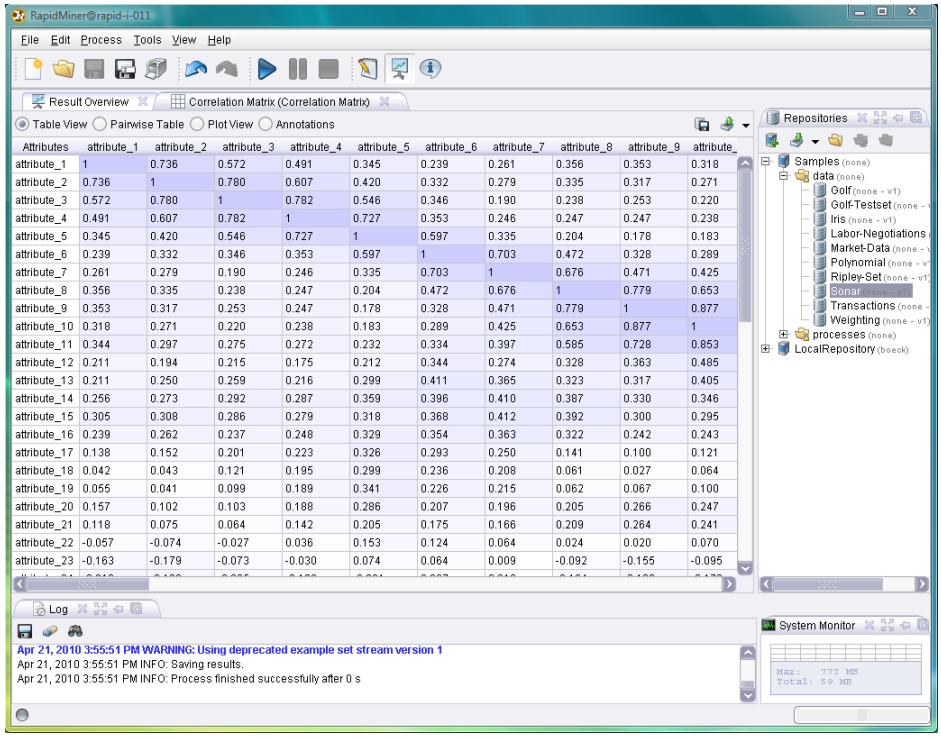
Abbildung 4.6: Farbkodierungen und alternierende Zeilenhintergründe erleichtern die Navigation innerhalb von Tabellen.

Diese Farbkodierung setzt sich auch in den Metadaten durch: Hier haben Attribute mit speziellen Rollen ebenfalls einen durchgängig hellgelben Hintergrund und die regulären Attribute alternierend hellblaue und weiße. Ganz anders kann dieses Farbschema, wie in Abbildung 4.7, jedoch für andere Objekte aussehen. Bei einer Korrelationsmatrix beispielsweise können auch einzelne Zellen eingefärbt sein: Je dunkler, desto stärker ist die Korrelation zwischen diesen Attributen .

Sortierung

Die meisten Tabellen können in RapidMiner mit einem simplen Klick sortiert werden. Bewegen Sie den Mauszeiger etwa in die Mitte der Spaltenüberschrift und klicken Sie die Überschrift an. Ein kleines Dreieck zeigt nun die Richtung der Sortierung an. Ein weiterer Klick ändert die Sortierrichtung und noch ein Klick würde die Sortierung wieder deaktivieren.

Sie können auch nach mehreren Spalten gleichzeitig sortieren, d.h. zunächst nach

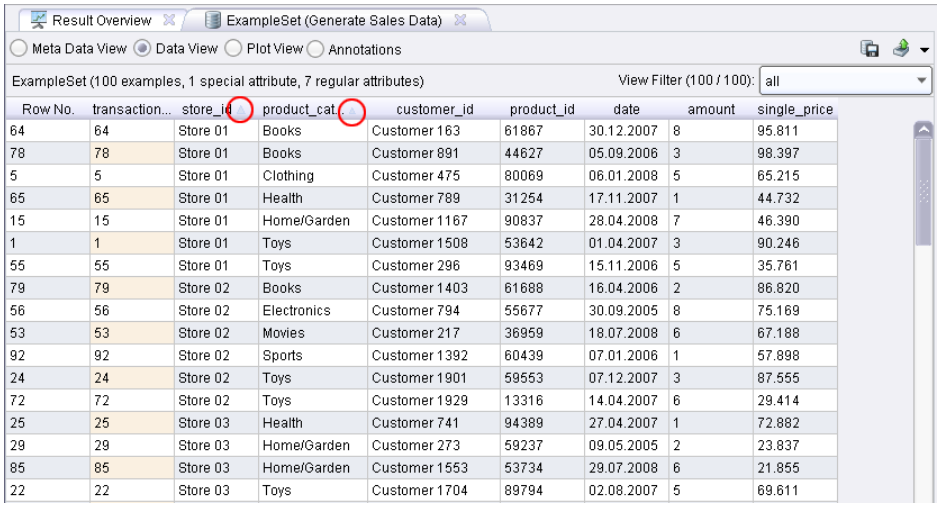


Attributes	attribute_1	attribute_2	attribute_3	attribute_4	attribute_5	attribute_6	attribute_7	attribute_8	attribute_9	attribute_
attribute_1	1	0.736	0.572	0.491	0.345	0.239	0.261	0.356	0.353	0.318
attribute_2	0.736	1	0.780	0.607	0.420	0.332	0.279	0.335	0.317	0.271
attribute_3	0.572	0.780	1	0.782	0.546	0.346	0.190	0.238	0.253	0.220
attribute_4	0.491	0.607	0.782	1	0.727	0.353	0.246	0.247	0.247	0.238
attribute_5	0.345	0.420	0.546	0.727	1	0.597	0.335	0.204	0.178	0.183
attribute_6	0.239	0.332	0.346	0.353	0.597	1	0.703	0.472	0.328	0.289
attribute_7	0.261	0.279	0.190	0.246	0.335	0.703	1	0.676	0.471	0.425
attribute_8	0.356	0.335	0.238	0.247	0.204	0.472	0.676	1	0.779	0.653
attribute_9	0.353	0.317	0.253	0.247	0.178	0.328	0.471	0.779	1	0.877
attribute_10	0.318	0.271	0.220	0.238	0.183	0.289	0.425	0.653	0.877	1
attribute_11	0.344	0.297	0.275	0.272	0.232	0.334	0.397	0.585	0.728	0.853
attribute_12	0.211	0.194	0.215	0.175	0.212	0.344	0.274	0.328	0.363	0.485
attribute_13	0.211	0.250	0.259	0.216	0.299	0.411	0.365	0.323	0.317	0.405
attribute_14	0.256	0.273	0.292	0.287	0.359	0.396	0.410	0.387	0.330	0.346
attribute_15	0.305	0.308	0.286	0.279	0.318	0.368	0.412	0.392	0.300	0.295
attribute_16	0.239	0.262	0.237	0.248	0.329	0.354	0.363	0.322	0.242	0.243
attribute_17	0.138	0.152	0.201	0.223	0.326	0.293	0.250	0.141	0.100	0.121
attribute_18	0.042	0.043	0.121	0.195	0.299	0.236	0.208	0.061	0.027	0.064
attribute_19	0.055	0.041	0.099	0.189	0.341	0.226	0.215	0.062	0.067	0.100
attribute_20	0.157	0.102	0.103	0.188	0.286	0.207	0.196	0.205	0.266	0.247
attribute_21	0.118	0.075	0.064	0.142	0.205	0.175	0.166	0.209	0.264	0.241
attribute_22	-0.057	-0.074	-0.027	0.036	0.153	0.124	0.064	0.024	0.020	0.070
attribute_23	-0.163	-0.179	-0.073	-0.030	0.074	0.064	0.009	-0.092	-0.155	-0.095

Abbildung 4.7: Tabellen in RapidMiner zeigen durch Farben häufig interessante Informationen an. In diesem Fall deuten dunklere Hintergründe auf stärkere Korrelationen zwischen Attributen hin.

einer Spalte sortieren und dann innerhalb dieser Sortierung noch nach bis zu zwei weiteren Spalten. Sortieren Sie hierzu zunächst auf die erste Spalte und sortieren Sie in die gewünschte Richtung. Drücken Sie nun die STRG-Taste und halten Sie diese gedrückt, während Sie weitere Spalten der Sortierung hinzufügen. Im folgenden Beispiel haben wir die Transaktionen zunächst nach der ID des Geschäfts und danach nach der Kategorie des Artikels sortiert. Die Reihenfolge der Spalten innerhalb dieser Sortierung wird durch verschieden große Dreiecke symbolisiert von groß nach klein (Abbildung 4.8).

Hinweis: Die Sortierung kann zeitaufwändig sein. Daher ist sie bei großen Tabellen deaktiviert, damit nicht versehentlich eine Sortierung gestartet wird und das Programm in dieser Zeit nicht benutzbar ist. Sie können den Schwellwert, ab dem die Sortierung deaktiviert wird, in den Einstellungen unter „Tools“ – „Preferences“ einstellen.



Result Overview | ExampleSet (Generate Sales Data)

Meta Data View | Data View | Plot View | Annotations

ExampleSet (100 examples, 1 special attribute, 7 regular attributes) View Filter (100 / 100): all

Row No.	transaction...	store_id	product_cat...	customer_id	product_id	date	amount	single_price
64	64	Store 01	Books	Customer 163	61867	30.12.2007	8	95.811
78	78	Store 01	Books	Customer 891	44627	05.09.2006	3	98.397
5	5	Store 01	Clothing	Customer 475	80069	06.01.2008	5	65.215
65	65	Store 01	Health	Customer 789	31254	17.11.2007	1	44.732
15	15	Store 01	Home/Garden	Customer 1167	90837	28.04.2008	7	46.390
1	1	Store 01	Toys	Customer 1508	53642	01.04.2007	3	90.246
55	55	Store 01	Toys	Customer 296	93469	15.11.2006	5	35.761
79	79	Store 02	Books	Customer 1403	61688	16.04.2006	2	86.820
56	56	Store 02	Electronics	Customer 794	55677	30.09.2005	8	75.169
53	53	Store 02	Movies	Customer 217	36959	18.07.2008	6	67.188
92	92	Store 02	Sports	Customer 1392	60439	07.01.2006	1	57.898
24	24	Store 02	Toys	Customer 1901	59553	07.12.2007	3	87.555
72	72	Store 02	Toys	Customer 1929	13316	14.04.2007	6	29.414
25	25	Store 03	Health	Customer 741	94389	27.04.2007	1	72.882
29	29	Store 03	Home/Garden	Customer 273	59237	09.05.2005	2	23.837
85	85	Store 03	Home/Garden	Customer 1553	53734	29.07.2008	6	21.855
22	22	Store 03	Toys	Customer 1704	89794	02.08.2007	5	69.611

Abbildung 4.8: In dieser Tabelle wurde zunächst nach dem Attribut „store_id“ aufsteigend sortiert und dann innerhalb der Store-ID-Blöcke ebenfalls aufsteigend nach der Produktkategorie.

Bewegen von Spalten

Sie können bei den meisten Tabellen die Reihenfolge der Spalten ändern, indem Sie auf die Spaltenüberschrift klicken und bei gedrückter Maustaste die Spalte an eine neue Position ziehen. Dies kann praktisch sein, wenn Sie die Inhalte zweier Spalten in umfangreichen Tabellen miteinander vergleichen wollen.

Anpassen von Spaltenbreiten

Sie können die Breite von Spalten anpassen, indem Sie den Mauszeiger über den Bereich zwischen zwei Spalten halten und bei gedrückter Maustaste die Breite der Spalte links von dem Trennbereich ändern. Alternativ können Sie auch einen Doppelklick auf diesen Zwischenraum durchführen, wodurch die Breite der Spalte links von dem Zwischenraum automatisch auf die notwendige Mindestgröße eingestellt wird. Zu guter Letzt können Sie während eines solchen Doppelklicks auf einen Spaltenzwischenraum auch noch die STRG-Taste gedrückt halten, wodurch die Größe aller Spalten automatisch angepasst wird.

Tip: Die Kombination von STRG und dem Doppelklick auf einen Spaltenzwischenraum im Bereich der Spaltenüberschriften sollten Sie sich merken zum Schnellen einstellen der Spaltenbreiten.

Aktionen im Kontextmenü

Sie können in den meisten Tabellen mit einem Rechtsklick auf eine Tabellenzelle ein Kontextmenü mit weiteren Aktionen öffnen. Im Einzelnen umfassen diese Aktionen:

1. *Select Row:* Auswahl einer Zeile,
2. *Select Column:* Auswahl einer Spalte,
3. *Fit Column Width:* Anpassen der Breite der ausgewählten Spalte,
4. *Fit all Column Widths:* Anpassen aller Spaltenbreiten,
5. *Equal Column Widths:* Verwendung einer gleichen Standardbreite für alle Spalten,
6. *Sort by Column (Ascending):* Aufsteigende Sortierung nach dieser Spalte,
7. *Sort by Column (Descending):* Absteigende Sortierung nach dieser Spalte,
8. *Add to Sorting Columns (Ascending):* Hinzufügen zu den Sortierspalten (aufsteigend),
9. *Add to Sorting Columns (Descending):* Hinzufügen zu den Sortierspalten (absteigend),
10. *Sort Columns by Names:* Neuanordnung der Spalten nach alphabetischer Sortierung der Spaltenüberschriften,
11. *Restore Column Order:* Wiederherstellung der ursprünglichen Spaltenanordnung.

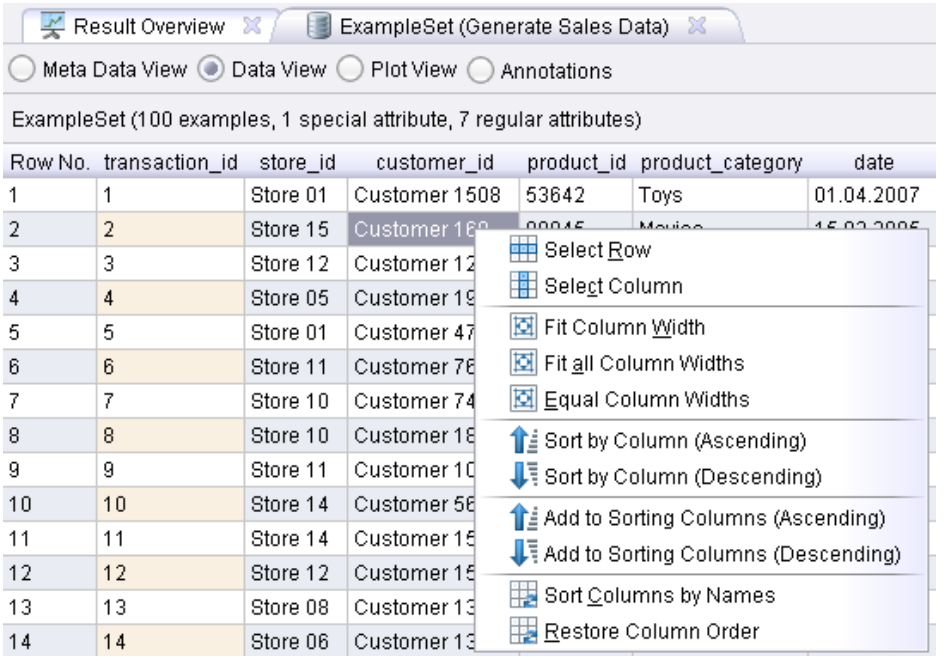

Result Overview
ExampleSet (Generate Sales Data)
Meta Data View
Data View
Plot View
Annotations
ExampleSet (100 examples, 1 special attribute, 7 regular attributes)
Row No. transaction_id store_id customer_id product_id product_category date
1 1 Store 01 Customer 1508 53642 Toys 01.04.2007
2 2 Store 15
3 3 Store 12
4 4 Store 05
5 5 Store 01
6 6 Store 11
7 7 Store 10
8 8 Store 10
9 9 Store 11
10 10 Store 14
11 11 Store 14
12 12 Store 12
13 13 Store 08
14 14 Store 06
Select Row
Select Column
Fit Column Width
Fit all Column Widths
Equal Column Widths
Sort by Column (Ascending)
Sort by Column (Descending)
Add to Sorting Columns (Ascending)
Add to Sorting Columns (Descending)
Sort Columns by Names
Restore Column Order

Abbildung 4.9: Aktionen wie die Auswahl von Zeilen oder Spalten, Sortieren der Inhalte nach Spalten oder die Anpassung von Spaltenbreiten stehen in einem Kontextmenü zur Verfügung.

Kopieren von Tabelleninhalten

Genau wie bei der Textansicht oben können Sie auch innerhalb von Tabellen einzelne Zellen mit der Maus markieren oder die vollständige Tabelle durch einen Klick in die Tabelle und mittels STRG + A. Zusätzlich stehen Ihnen im Kontextmenü noch Aktionen zur Verfügung, um ganze Zeilen oder Spalten zu markieren. Danach können Sie den ausgewählten Bereich mittels STRG + C in die Zwischenablage kopieren und in andere Applikationen einfügen. Beachten Sie bitte, dass hierbei die Tabellenstruktur erhalten bleibt, wenn Sie beispielsweise in Anwendungen wie Microsoft Excel einfügen, die ihrerseits tabellarische Daten unterstützen.

4.4.3 Plotter

Eine der stärksten Eigenschaften von RapidMiner sind die zahlreichen Visualisierungsverfahren sowohl für Daten und andere Tabellen wie auch für Modellierungen. Solche Visualisierungen werden dem Analysten typischerweise in der „Plot View“ angeboten.

Konfiguration von Plottern

Der Aufbau aller Plotter in RapidMiner ist prinzipiell gleich. Auf der linken Seite befindet sich ein Konfigurationsbereich, der aus mehreren wiederkehrenden Elementen besteht:

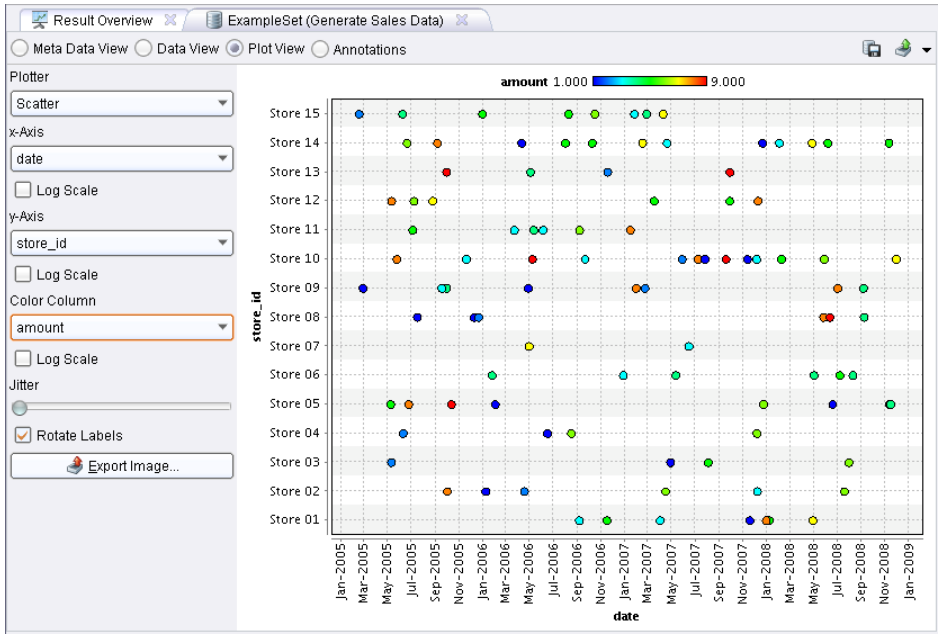


Abbildung 4.10: Visualisierung eines Datensatzes und die Plotter-Konfiguration auf der linken Seite.

Die wichtigste Einstellung ist ganz oben zu finden und entspricht dem Typ der Visualisierung. Es stehen mehr als 30 verschiedene 2D-, 3D- und auch hochdimensionale Visualisierungsverfahren zur Darstellung Ihrer Daten und Ergebnisse zur Verfügung. Im Bild oben sehen Sie einen Plot des Typs „Scatter“. Je nach

Auswahl des Plotter-Typs ändern sich alle weiteren Einstellungsfelder. Bei einem Scatter-Plot beispielsweise geben Sie die Attribute für die x-Achse und für die y-Achse an und können noch ein drittes Attribut zur Einfärbung der Punkte verwenden. Speziell für den Scatter-Plot gibt es noch weitere Möglichkeiten wie beispielsweise die Angaben, ob die Achsen logarithmisch skaliert werden sollen.

Tip: Speziell für Datensätze, welche nicht nur Zahlen sondern auch nominale Werte beinhalten, ist die Funktion „Jitter“ sehr hilfreich. Hiermit geben Sie an, ob und wie weit die Punkte von ihrer ursprünglichen Position weg in eine zufällige Richtung bewegt werden sollen. Damit können Sie Punkte, die ansonsten durch andere Punkte überdeckt werden würden, leicht sichtbar machen.

Viele Plotter erlauben darüber hinaus auch noch weitere Konfigurationen der Darstellung, beispielsweise ob die Beschriftung an der x-Achse rotiert werden soll, so dass auch lange Texte noch lesbar bleiben. Probieren Sie einfach ein wenig mit den Einstellungen und den verschiedenen Möglichkeiten herum, Sie werden schon bald mit den zahlreichen Möglichkeiten zur Visualisierung vertraut sein.

Tip: Die verwendeten Farben können Sie übrigens in den Einstellungen unter „Tools“ – „Preferences“ ändern.

Änderung des Plotter-Typs

Die Auswahl des Plotter-Typs definiert maßgeblich, welche Parameter Sie einstellen können. In Abbildung 4.11 sehen Sie ein Beispiel für einen Plotter des Typs „Bars Stacked“. Statt der verschiedenen Achsen stellen Sie nun Attribute ein, nachdem die Daten gruppiert werden sollen (hier: „store_id“) und welches Attribut zur Definition der Stacks verwendet werden soll (hier: „product_category“). Die Höhe der Balken entspricht dann der Summe (hier: „Aggregation“ steht auf „Sum“) des als Value Column definierten Attributes (hier: „amount“).

Berechnung von Visualisierungen

Zu guter Letzt soll an dieser Stelle noch erwähnt werden, dass es noch Visualisierungen gibt, die ihrerseits so aufwändig sind, dass Sie eigens berechnet werden müssen. Solche Visualisierungen, wie beispielsweise eine Self-Organizing-Map (SOM) bieten dann einen Knopf namens „Calculate“, mit dem die Berechnung und in Abbilung 4.12 dargestellte Visualisierung gestartet werden kann.

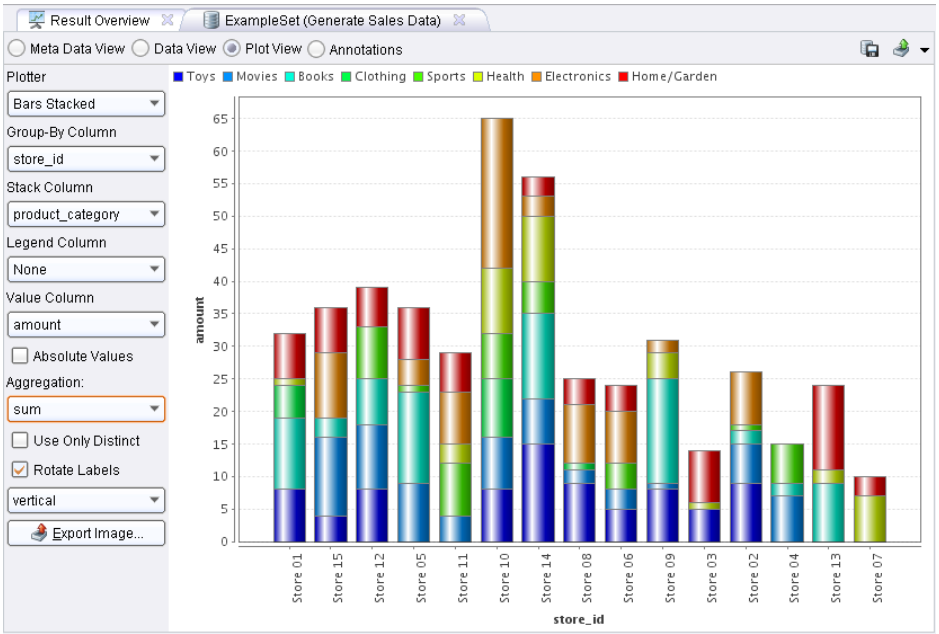


Abbildung 4.11: Änderung der Plotter-Konfiguration in Abhängigkeit von dem Plotter-Typ.

4.4.4 Graphen

Graphen sind eine weitere Darstellungsform, welche relativ häufig in RapidMiner zu finden sind. Prinzipiell verstehen wir hierunter alle Visualisierungen, welche Knoten und ihre Beziehungen zeigen. Das können Knoten innerhalb eines hierarchischen Clusterings sein oder auch wie in Abbildung 4.13 die Knoten eines Entscheidungsbaums.

Graphen wie der des obigen Entscheidungsbaums werden zumeist als „Graph View“ bezeichnet und stehen unter diesem Namen zur Verfügung.

Zooming

Sie können mittels des Mausrads, sofern vorhanden, in den Graphen hinein und aus einem Graphen heraus zoomen. Alternativ stehen Ihnen im Konfigurationsbereich oben links auch zwei Schaltflächen zur Verfügung, um den Zoom-Level ihres Graphen zu vergrößern und zu verkleinern.

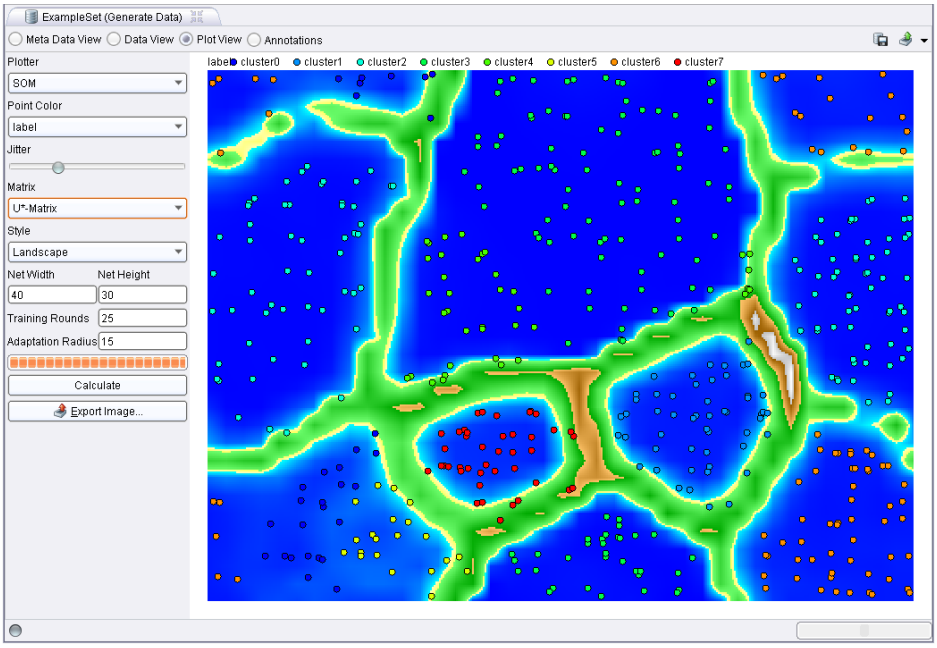


Abbildung 4.12: Aufwändige Visualisierungen wie beispielsweise SOMs bieten einen Knopf „Calculate“, um die Berechnung zu starten. Der Fortschritt wird mittels eines Balkens angezeigt.

Modus

Es stehen zwei grundlegende Navigationsweisen im Graphen zur Verfügung, die auch als Modus bezeichnet werden:

1. *Verschieben*: Der Modus zum Verschieben des Graphen wird durch die linke Schaltfläche in der Modus-Box ausgewählt. In diesem Fall können Sie mit gedrückter linker Maustaste den Ausschnitt des Graphen verschieben, um sich so verschiedene Bereiche des Graphen detailliert ansehen zu können.

2. *Auswählen*: Der Modus zum Auswählen einzelner Knoten wird durch die rechte Schaltfläche in der Modus-Box ausgewählt. Nun können Sie einzelne Knoten mittels Klicks auswählen oder mit gedrückter Maustaste in einen freien Bereich einen Auswahlrahmen für mehrere Knoten zugleich definieren. Mittelst gedrückter SHIFT-Taste können Sie einzelne Knoten der Auswahl hinzufügen oder diese von der Auswahl ausschließen. Gerade

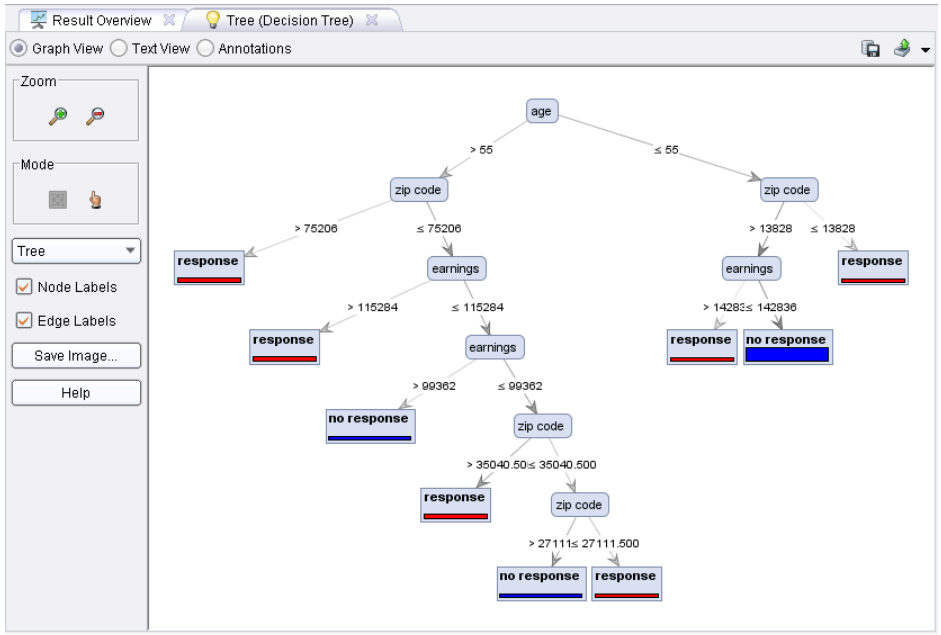


Abbildung 4.13: Ein Entscheidungsbaum in einer Graphansicht.

ausgewählte Knoten können mit gedrückter Maustaste verschoben werden.

Weitere Hinweise zu der Bedienung von Graphen in diesen beiden Modi finden Sie im Hilfe-Dialog, der angezeigt wird, wenn Sie auf den Knopf „Help“ im Konfigurationsbereich des Graphen klicken.

Weitere Einstellungen

Sie können einstellen, ob die Beschriftungen für Knoten und Kanten angezeigt werden sollen oder nicht. Die wichtigste Einstellung, nicht unbedingt für Bäume aber für andere Graphen, ist die Wahl eines passenden Layouts, was in der Auswahlbox direkt unterhalb der Modusbox geschehen kann. Die verschiedenen Algorithmen haben unterschiedliche Stärken und Schwächen und Sie müssen üblicherweise ausprobieren, welche Darstellung für den vorliegenden Graphen das beste Ergebnis liefert.

4.4.5 Spezielle Ansichten

Neben den oben beschriebenen Views Text, Tabelle, Plotter und Graph gibt es vereinzelt auch weitere Darstellungskomponenten, die jedoch seltener vorkommen und selbsterklärend sein sollten. So gibt es beispielsweise für Frequent Itemsets noch eine eigene Art von Tabelle oder ein spezieller Graph für die zugehörigen Assoziationsregeln.

4.5 Result Overview

Wir haben eingangs bereits den Result Overview bemerkt, welcher als eine Art Platzhalter stets an der Stelle zu finden ist, an der auch die übrigen Resultate angezeigt werden:

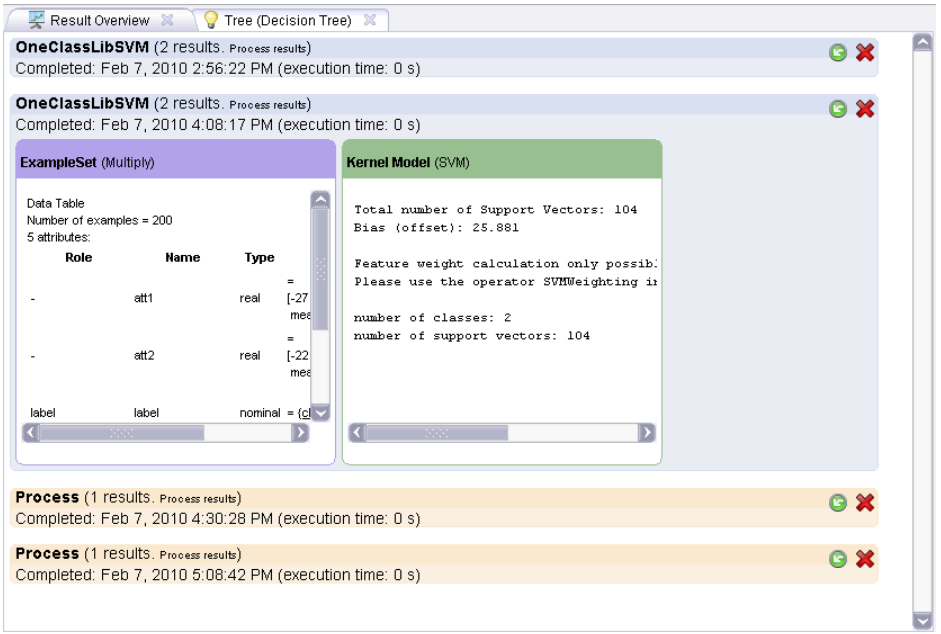


Abbildung 4.14: Die Result Overview zeigt die Ergebnisse der letzten Analyseprozesse an.

Die Result Overview dient als kompakte Übersicht über alle Prozessausführungen der aktueller RapidMiner-Sitzung. Jeder zweizeilige Eintrag besteht aus dem Na-

men des Prozesses, der Anzahl der Ergebnisse sowie Informationen darüber, wann der Prozess beendet wurde und wie lange er lief. Jeweils blockweise abwechselnd sind die Ergebnisse des gleichen Prozesses eingefärbt.

Sie können durch einen Klick auf einen Eintrag eine Detailansicht der Ergebnisse einsehen. Im Fall oben besteht das Ergebnis aus einem Example Set und einem SVM-Modell. Ein weitere Klick auf den Eintrag schließt diesen wieder. Natürlich können Sie auch mehrere Einträge gleichzeitig öffnen und so die Ergebnisse bequem vergleichen.

Für jeden Eintrag stehen oben rechts zwei Aktionen zur Verfügung, nämlich

1. den Prozess, der zu einem Eintrag gehört, in dieser Form wieder herzustellen und

2. den Eintrag aus der Result Overview zu löschen.

Darüber hinaus steht Ihnen in den Kontextmenüs der Overview und der einzelnen Beiträge auch noch die Option zur Verfügung, die vollständige Overview zu löschen.

Hinweis: Wenn Sie die Result Overview schließen möchten, warnt RapidMiner Sie mit einem Hinweis darauf, dass in dieser Perspektive keine Ergebnisse mehr angezeigt werden. Wir empfehlen also dringend, die Result Overview nicht zu schließen beziehungsweise mindestens in einer Perspektive einen Result Overview geöffnet zu lassen.

5 Verwaltung von Daten: Das Repository

Tabellen, Datenbanken, Textsammlungen, Logdateien, Webseiten, Messwerte – dies und Ähnliches steht am Anfang jedes Data Mining Prozesses. Daten werden aufbereitet, umgewandelt, zusammengeführt, und am Ende erhalten Sie neue oder anders repräsentierte Daten, Modelle oder Berichte. In diesem Kapitel erfahren Sie, wie Sie all diese Objekte mit RapidMiner handhaben.

5.1 Das RapidMiner Repository

Sobald Ihre Sammlung von Prozessen und den mit ihnen assoziierten Dateien eine gewisse Größe übersteigt, werden Sie feststellen, dass es ratsam ist, diese auf eine konsistente und strukturierte Art und Weise zu organisieren. Eine Möglichkeit ist die Organisation von Projekten auf Dateiebene. Dateien werden zu Projekten gruppiert und jeweils ein Verzeichnis für Ausgangsdaten, Zwischenergebnisse, Berichte, etc. angelegt.

Während das Anlegen aufgeräumter Projektstrukturen eine sinnvolle Sache ist, ist die Verwendung des normalen Dateisystems in den seltensten Fällen angeraten und für die Bedürfnisse einer Data Mining Lösung kaum ausreichend. Verschiedene Gründe wie Vertraulichkeit oder begrenzter Speicherplatz können das Ablegen von Dateien auf dem lokalen Rechner unmöglich machen. Soll ein auf dem lokalen Rechner erstellter Prozess auf einem entfernten Server ausgeführt werden, erfordert dies manuelle Eingriffe wie das Kopieren des Prozesses und das Anpassen von Pfaden. Kollaboratives Erstellen von Prozessen, Bearbeiten von Daten und Auswerten von Ergebnissen erfordert eine externe Rechte- und

Versionsverwaltung. In unterschiedlichen Formaten abgelegte Dateien erfordern die korrekte Einstellung von Parametern wie Trennzeichen und Kodierung bei jedem neuen Einladen. Zwischenergebnisse und Prozessvarianten wachsen schnell zu einer beachtlichen Anzahl an, so dass man leicht die Übersicht verlieren kann. Das Einladen und Betrachten von Daten zwecks Wiedergewinnung der Übersicht erfordert einen unter Umständen langwierigen Einladevorgang oder sogar den Start einer externen Applikation. Annotationen von Dateien, die dies erleichtern können, werden von normalen Dateisystemen nicht unterstützt.

RapidMiners Antwort auf all diese Probleme ist das Repository, das alle Daten und Prozesse aufnimmt. Zwar können Daten auch von außerhalb des Repositorys in Prozesse einfließen, was z.B. für die Ausführung von ETL Prozessen nötig ist, die Verwendung des Repositorys bietet jedoch eine Reihe von Vorteilen, die Sie nicht werden missen wollen:

- Daten, Prozesse, Ergebnisse und Berichte werden an relativ zueinander angegebenen Orten in einem für den Nutzer transparenten Mechanismus abgespeichert.
- Das Öffnen oder Einladen der Dateien erfordert keine weiteren Einstellungen. Daten können durch einen einzelnen Klick geöffnet, betrachtet oder in den Prozess eingebaut werden. Eine Übersicht über die abgespeicherten Daten, ihre Eigenschaften und von Ihnen selbst vergebene Bemerkungen bekommen Sie jederzeit ohne die Datei einzeln öffnen zu müssen.
- Alle Ein- und Ausgabedaten sowie Zwischenergebnisse werden mit Metainformationen annotiert. Dies garantiert Konsistenz und Integrität Ihrer Daten und erlaubt die Validierung von Prozessen zur Entwicklungszeit sowie das Bereitstellen von kontextsensitiven Assistenten.

Das Repository kann entweder auf einem lokalen oder geteilten Dateisystem liegen oder durch den externen RapidMiner Analyseserver namens RapidAnalytics bereitgestellt werden. Die folgende Abbildung zeigt den Repository View, der den Inhalt des Repositorys darstellt. RapidMiner stellt einen Satz von Beispielprozessen und -daten zur Verfügung, die Sie im initial angelegten Repository finden. Einige von diesen sind in der Abbildung 5.1 zu sehen.

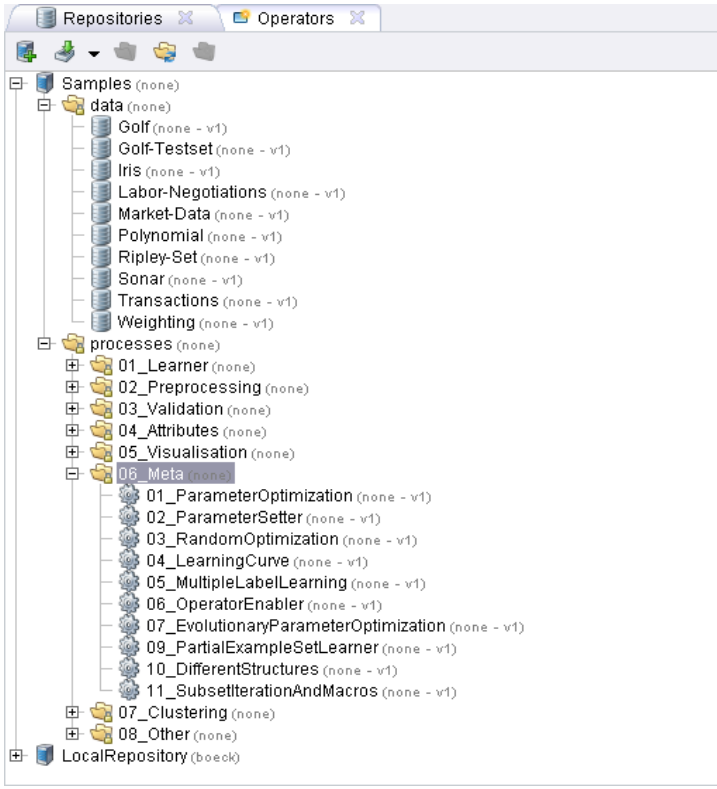


Abbildung 5.1: Der Repository View mit einem geöffneten Beispielverzeichnis.

5.1.1 Ein neues Repository anlegen

Um das Repository benutzen zu können, müssen Sie zunächst eine solches erstellen. RapidMiner fordert Sie auf, dies zu tun, wenn es zum ersten Mal gestartet wird. Später können Sie weitere Repositories hinzufügen, indem Sie die erste Schaltfläche in der Werkzeugleiste der Repository View benutzen. Die folgenden Abbildungen zeigen den einfachen Ablauf. Sofern Sie nicht über den Analyseserver von RapidAnalytics verfügen, wählen Sie die erste Option, um ein lokales Repository anzulegen und wählen Sie dann Next. Vergeben Sie nun einen Namen für Ihr Repository und wählen Sie ein Verzeichnis, in dem es angelegt werden soll. Schließen Sie den Dialog mit Finish ab. Sie können Ihr Repository nun verwenden.

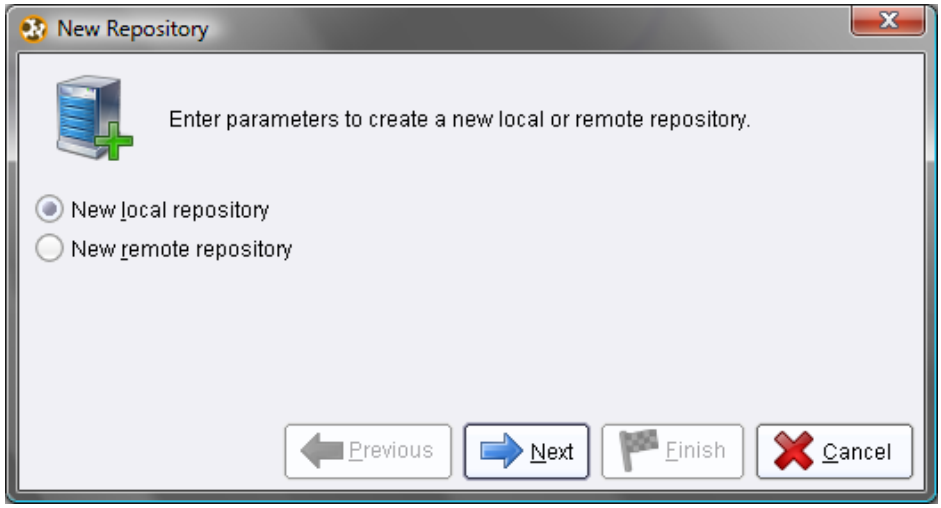


Abbildung 5.2: Sie können ein Repository auf einem gemeinsam genutzten Analyseserver RapidAnalytics nutzen oder ein lokales Repository auswählen.

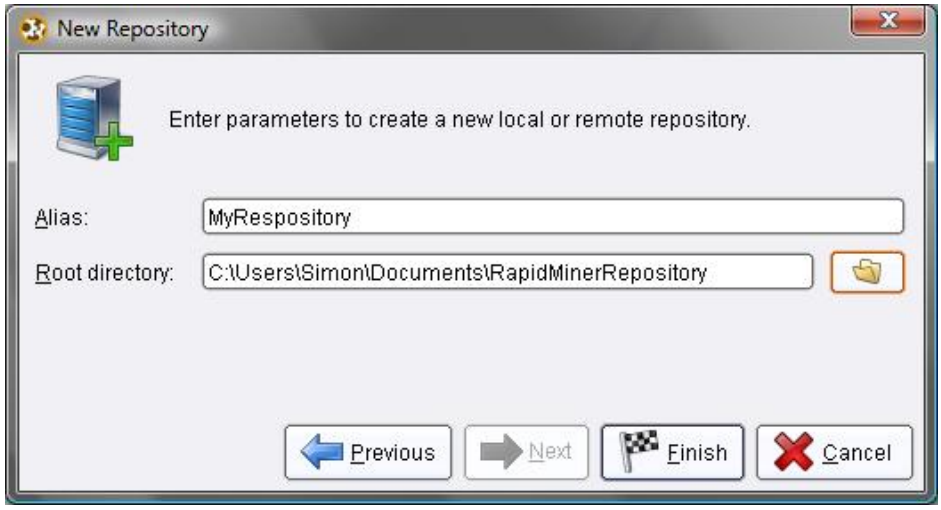


Abbildung 5.3: RapidMiner erfragt Namen und Verzeichnis für ein neu angelegtes lokales Repository. auswählen.

5.2 Das Repository verwenden

Es bietet sich an, für Projekte eine einheitliche Verzeichnisstruktur zu verwenden, beispielsweise einen Projektordner mit dem Namen des Projekts, und jeweils

einen Ordner für Prozesse, Eingabedaten und Ergebnisse. Dieser Struktur folgen alle Beispiele in diesem Buch. Verzeichnisse erstellen können Sie mit Hilfe des Kontextmenüs im Repository View oder mit Hilfe der Schaltfläche in der Werkzeugleiste oben in diesem View.

5.2.1 Prozesse und relative Repositoryangaben

Bevor wir in den nächsten Abschnitten diskutieren, wie Sie Daten und Prozesse im Repository ablegen können und wieder auf diese zugreifen, wollen wir zunächst einige grundsätzliche Hinweise zur Referenzierung dieser Objekte innerhalb des Repositorys geben. Prozesse können Sie im Repository abspeichern, indem Sie im Kontextmenü den Eintrag „Store Process“ wählen oder indem Sie den entsprechenden Eintrag im „File“-Menü wählen. Es öffnet sich im letzteren Fall noch der Repository Browser, in dem Sie den Ort zum Abspeichern des Prozesses angeben können. Nachdem ein Prozess im Repository abgespeichert ist, werden alle Referenzen auf Repositoryeinträge, die als Parameter von Operatoren gesetzt sind, relativ zum Ort des Prozesses aufgelöst. Was heißt das? Einträge im Repository werden nach folgendem Schema bezeichnet:

//RepositoryName/Ordner/Unterordner/Datei

Die doppelten Schrägstriche am Beginn zeigen an, dass zunächst der Name eines Repositorys folgt. Anschließend folgen weitere Ordnernamen und abschließend ein Dateiname. Wir nennen solche Angaben *absolut*. Der Angabe

/Ordner/Unterordner/Datei

fehlt die führende Repositorybezeichnung. Diese Angabe ist daher *Repository-relativ*. Sie bezieht sich auf den angegebenen Ordner im selben Repository, in dem der Prozess liegt, in dem diese Angabe verwendet wird. Der führende Schrägstrich kennzeichnet hier eine absolute Pfadangabe. Fehlt auch dieser, wird die Angabe *relativ* aufgelöst:

../RelativerOrdner/Datei

bezeichnet beispielsweise eine Datei im Ordner „RelativerOrdner“, den wir erreichen,

indem wir von demjenigen Ordner, der den aktuellen Prozess enthält, ein Ver-

zeichnis

nach oben wandern („..“) und dort den Ordner „RelativerOrdner“ suchen. Befindet sich der Prozess also beispielsweise in der Datei

//MeinRepository/ProjektA/Prozesse/ProzessB,

führt diese Angabe nach

//MeinRepository/ProjektA/RelativerOrdner/Datei.

Hinweis: Die Beschreibungen oben klingen wahrscheinlich komplizierter als sie in der Praxis wirklich sind. Solange Sie als allererstes für jeden neuen Prozess einen Platz innerhalb des Repositories definieren und danach einfach für alle Operatorparameter, die einen Eintrag im Repository erfordern den Repository Browser verwenden, achtet RapidMiner vollständig automatisch darauf, nach Möglichkeit immer relative Angaben zu verwenden. Dies erleichtert insbesondere Restrukturierungen des Repositorys und Kopien für andere Anwender, was bei absoluten Angaben schwierig wäre.

5.2.2 Daten und Objekte in das Repository importieren

Es gibt zahlreiche Möglichkeiten, Daten und andere Objekte wie Modelle in das Repository einzupflegen. Wir beschreiben an dieser Stelle die wichtigsten.

ExampleSets mit Wizards importieren

Haben Sie Daten in einem bestimmten Format vorliegen und wollen Sie diese in einem RapidMiner-Prozess benutzen, stehen Ihnen für viele Dateiformate und Datenbanken sogenannte *Wizards* zur Verfügung. Ein Wizard ist ein Dialog, der Sie Schritt für Schritt durch den Einladeprozess führt. Allen Wizards ist gemeinsam, dass Sie bestimmte Metadaten wie Attributtypen, Wertebereiche und Rollen für die einzelnen Spalten vergeben können. Im oberen Bereich des Repositorys finden Sie ein Icon, welches für den ausgewählten Dateityp den passenden Wizard startet. Dieselbe Aktion finden Sie auch im „File“-Menü von RapidMiner. Schließlich gibt es auch noch eine besonders einfache Weise für den Import von Dateien: Ziehen Sie die zu importierende Datei einfach bei gedrückter Maustaste in das Repository. Sofern möglich, wird daraufhin ein passender Wizard gestartet.

Der Operator „Store“

Haben Sie einen ETL-Prozess oder einen anderen Prozess, dessen Ergebnis Sie im Repository abspeichern möchten, können Sie dieses tun, indem Sie den Operator „Store“ in Ihren Prozess einbauen.

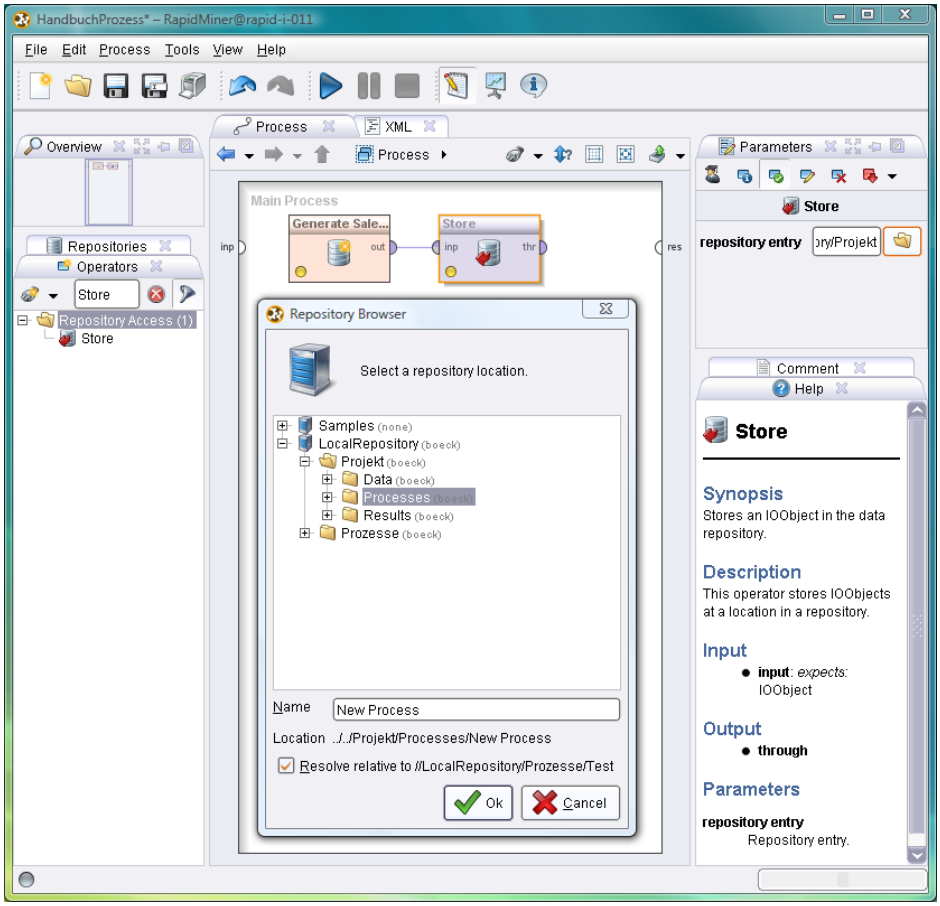


Abbildung 5.4: Der Operator „Store“ kann verwendet werden, um beliebige Daten und Objekte im Repository zu speichern. Der Dialog zeigt den Repository Browser, um den Speicherort festzulegen, und erscheint bei Klick auf den „Verzeichnis“-Knopf in den Parametern des Operators.

Der Beispielsprozess in dieser Abbildung generiert mit Hilfe des Operators „Generate Data“ einen Datensatz, der ins Repository gespeichert werden soll. Der „Sto-

re“-Operator hat nur einen einzigen Parameter, „repository_location“. Wählen Sie die Schaltfläche mit dem Ordner neben diesem Parameter, erhalten Sie einen Dialog, in dem Sie zunächst einen Ordner im Repository und dann einen Namen für den Datensatz vergeben können. Führen Sie den Prozess aus, werden Sie sehen, dass Sie einen neuen Eintrag im Repository erhalten, der den generierten Datensatz enthält. Der Store-Operator ist damit insbesondere für Prozesse der Datenintegration und –transformation sinnvoll, die automatisch oder regelmäßig durchgeführt werden sollen, beispielsweise im Rahmen des Process Schedulers des Servers RapidAnalytics. Für eine einmalige und eher interaktive Integration von Daten ist sicher die oben beschriebene Verwendung der Wizards der häufiger verwendete Weg.

Hinweis: Sie können nicht nur Datensätze, sondern auch Modelle und alle anderen RapidMiner-Objekte mit dem Store-Operator verbinden. Damit können Sie auch beliebige Ergebnisse in Ihrem Repository speichern.

Import anderer Formate mittels Operatoren

Das Repository speichert Datensätze in einem Format ab, das alle von RapidMiner benötigten Daten und Metadaten enthält. Ihre Daten werden zu Beginn vermutlich in einem anderen Format vorliegen: CSV, Excel, SQL Datenbanken, etc. Wie oben beschrieben, können Sie diese Dateien in Ihr Repository überführen. RapidMiner kann jedoch auch zahlreiche andere Formate innerhalb von Prozessen importieren. Operatoren dazu finden Sie in der Gruppe „Import“. Bei der Benutzung dieser Operatoren ist jedoch Vorsicht geboten: Metadaten stehen für diese Operatoren nicht garantiert zur Verfügung, was beispielsweise dazu führen kann, dass Prozesse, die von der Existenz bestimmter Attributwerte ausgehen, mögliche Fehler erst zur Laufzeit des Prozesses bemerken. Dennoch ist die Verwendung dieser Dateiformate mitunter nicht vermeidbar, z.B. für die regelmäßige Ausführung von ETL-Prozessen. Das Ziel dieser Prozesse sollte es jedoch sein, die Daten mit einem nachfolgenden Store-Operator in das Repository zu überführen, so dass sie von den nachfolgenden eigentlichen Analyseprozessen verwendet werden können.

Die Operatoren der „Import“-Gruppe haben zahlreiche auf das jeweilige Format zugeschnittene Parameter. Deren Beschreibung entnehmen Sie bitte der jeweiligen Operatordokumentation.

Objekte aus der Ergebnis- oder Prozessansicht abspeichern

Nachdem Sie einen Prozess ausgeführt haben, wird Ihnen in der Grundeinstellung die Results-Perspektive mitsamt dem gleichnamigen Reiter präsentiert. In dessen Werkzeugleiste befindet sich auf der rechten Seite eine Schaltfläche, mit der Sie das aktuell gewählte Ergebnis im Repository abspeichern können. Auch hier erscheint ein Dialog, mit dem Sie einen Ordner und einen Namen auswählen können.

Enthält Ihr Prozess Zwischenergebnisse, die in der Results-Perspektive nicht (mehr) angezeigt werden, können Sie diese auch vom Process View aus abspeichern. Klicken Sie dazu mit der rechten Maustaste auf einen Port, an dem Daten anliegen. Dies ist an den Ausgangsports aller Operatoren, die bereits ausgeführt wurden, der Fall. Sie erkennen dies an der dunkleren Farbe und an einem entsprechenden Eintrag in der Kontexthilfe. Hier wählen Sie den Menüeintrag „Store in Repository“, um das Objekt abzuspeichern. Bitte beachten Sie jedoch, dass die Daten an den Ports mit der Zeit wieder freigegeben werden können, um Speicher zu sparen, und daher nicht garantiert und beliebig lange an den Ports anlegen. Vergleichen Sie hierzu bitte auch die Erläuterungen im vorigen Kapitel.

5.2.3 Zugriff und Verwaltung des Repositories

Haben Sie Ihre Daten einmal ins Repository eingepflegt, können Sie sie unter Verwendung des Retrieve-Operators in Ihren Prozessen verwenden. Sie können den Operator wie gewohnt aus dem Operators View in den Prozess ziehen und dort den Parameter zum Repository-Eintrag definieren. Es geht jedoch noch einfacher: Ziehen Sie einfach einen Eintrag im Repository, zum Beispiel einen Datensatz mit der Maus auf den Process View. Hier wird nun automatisch ein fertig konfigurierter Operator mit einer Referenz auf diesen Eintrag eingefügt. Handelt es sich bei den Eintrag um ein Objekt, wird ein neuer Operator vom Typ „Retrieve“ erzeugt und entsprechend konfiguriert. Handelt es sich bei dem Repository-Eintrag jedoch um einen Prozess, so wird ein neuer Operator vom Typ „Execute Process“ angelegt und dessen Parameter verweist automatisch auf den gewählten Prozess aus dem Repository.

Mit einem Rechtsklick auf Einträge im Repsitory erhalten Sie weitere Möglichkeiten, um auf das Repository zuzugreifen, die Sie von der Dateiverwaltung Ihres

Rechners kennen. Diese Aktionen sind auch über die Werkzeugleiste des Repository Views verfügbar. Weitestgehend sind diese Aktionen selbsterklärend:

1. *Store Process here:* speichert den aktuellen Prozess an den angegebenen Ort,
2. *Rename:* Benennt den Eintrag oder das Verzeichnis um,
3. *Create Folder:* Legt ein neues Verzeichnis an dieser Stelle an,
4. *Delete:* Löscht den gewählten Repository-Eintrag oder Verzeichnis,
5. *Copy:* Kopiert den gewählten Eintrag zum späteren Einfügen an anderen Stellen,
6. *Paste:* Kopiert einen zuvor kopierten Eintrag an diese Stelle,
7. *Copy Location to Clipboard:* Kopiert einen eindeutigen Bezeichner für diesen Eintrag in die Ablage, so dass Sie diese als Parameter für Operatoren, in Web Interfaces o.ä. nutzen können,
8. *Open Process:* Haben Sie einen Prozess ausgewählt, wird der aktuelle Prozess geschlossen und der gewählte geladen,
9. *Refresh:* Wenn das Repository auf einem gemeinsam genutzten Dateisystem liegt oder Sie den RapidMiner Analyseserver RapidAnalytics verwenden, so dass Daten zeitgleich von anderen Benutzern verändert werden können, können Sie hiermit die Ansicht des Repositorys auffrischen.

5.2.4 Der Prozesskontext

Wir haben schon zuvor die Output-Ports des Prozesses am rechten Rand des Process View verwendet, beispielsweise um die Ergebnisse des Prozesses in der Result-Perspektive sichtbar zu machen. Zusätzlich zu den Output-Ports des Prozesses gibt es auch noch Input-Ports, die Sie am linken Rand des Process View finden. Diese haben wir bisher nie verbunden. In der Grundeinstellung ist dies auch – zumindest für die Quellen – nicht sinnvoll, denn der Prozess selbst besitzt dann keine Eingabe. Die Verbindung der inneren Senken hat jedoch einen Effekt: Alle Objekte, die am Ende des Prozesses an einer Senke ankommen, werden in der Result-Perspektive als Ergebnis des Prozesses präsentiert.

Diese Input- und Output-Ports des Prozesses haben jedoch eine weitere Funktion. Ein typischer Prozess beginnt mit einer Reihe von Retrieve-Operatoren, auf die eine Reihe von verarbeitenden Operatoren folgen, und endet mit einer Reihe von Store-Operatoren. Das Erzeugen dieser Operatoren können Sie sich sparen, indem Sie den Context View benutzen, den Sie im „View“-Menü finden. Abbildung 5.5 zeigt diesen Context View.

Im Context View haben Sie die Möglichkeit, an die Eingabeports Daten aus einem Repository anzulegen und Ausgaben zurück ins Repository zu schreiben. Für jeden Port können Sie eine solche Angabe machen. Dies hat zwei Vorteile:

- Sie können sich die Operatoren für Retrieve und Store sparen und Ihr Prozess wird hierdurch oftmals etwas übersichtlicher.
- Die Verwendung des Kontextes ist weiterhin praktisch, um Prozesse zu testen, die mittels des Operators „Execute Process“ eingebunden werden sollen: Die Daten, die an diesem Operator anliegen, überschreiben die im Prozesskontext definierten Werte.

5.3 Daten und Metadaten

Außer den eigentlichen Daten speichert RapidMiner noch andere Informationen im Repository: Daten *über* die Daten, sogenannte Metadaten. Für jeden Typ von Objekten stehen solche Metadaten zur Verfügung, besonders sinnvoll eingesetzt werden können Sie aber insbesondere für Modelle und Datensätze. Die für Datensätze gespeicherten Metainformationen umfassen beispielsweise:

- die Anzahl der Beispiele,
- die Anzahl der Attribute,
- die Typen, Namen und Rollen der Attribute,
- die Wertebereiche der Attribute beziehungsweise einige grundlegende Statistiken,
- sowie die Anzahl der fehlenden Werte pro Attribut.

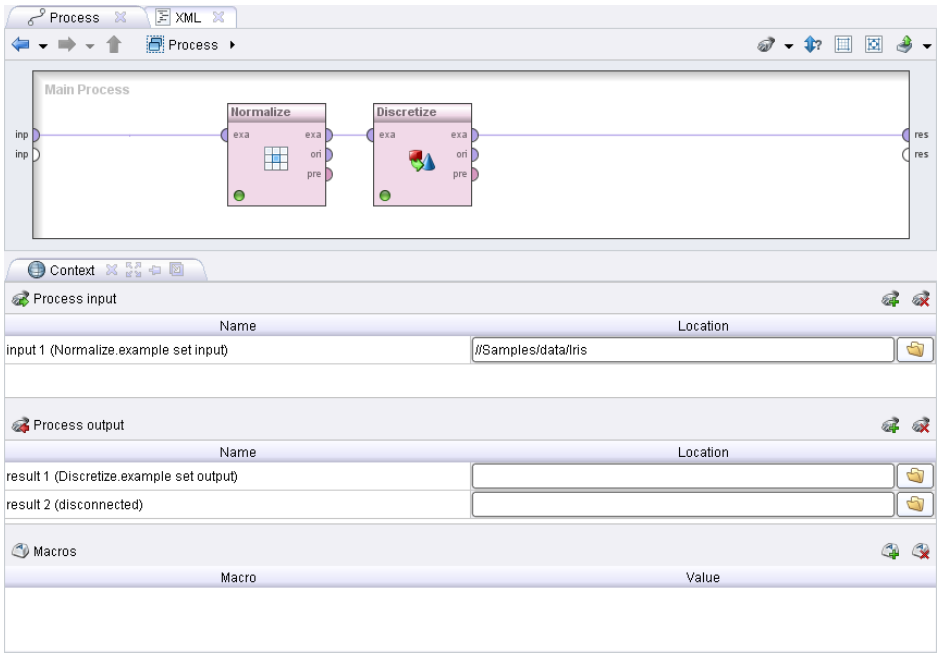


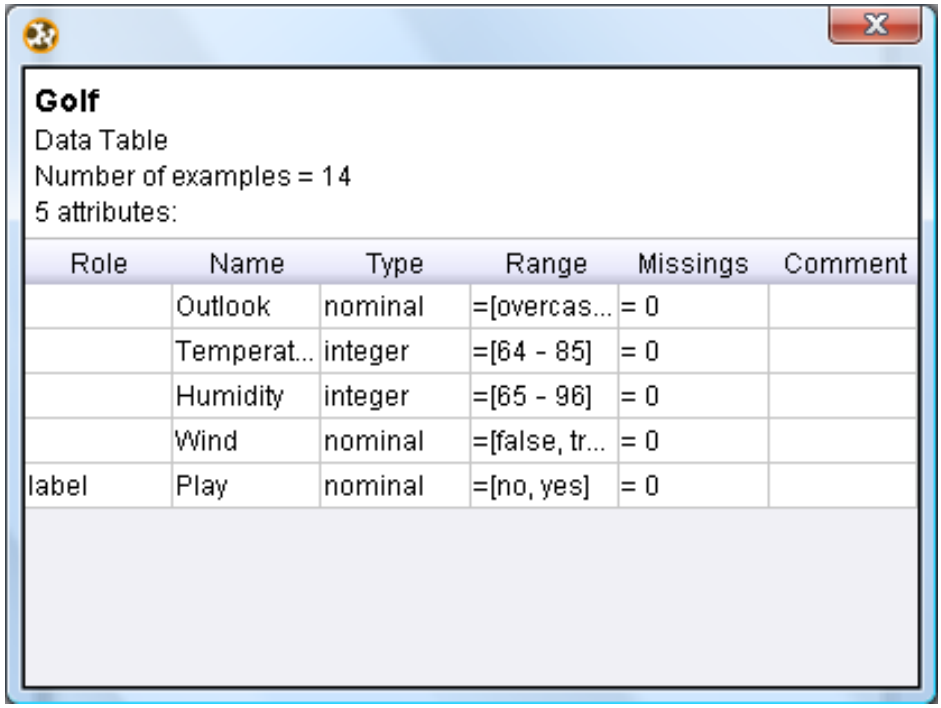
Abbildung 5.5: Der Prozesskontext. Bei „Input“ geben Sie Repositoryeinträge an, die als Eingabe des Prozesses dienen sollen und an Input-Ports des Prozesses angelegt werden. Bei „Output“ geben Sie an, wohin die Ergebnisse im Repository abgespeichert werden sollen.

Diese Informationen sind im Repository einsehbar, ohne den Datensatz zuvor einzuladen, was je nach Größe einige Zeit dauern kann. Bewegen Sie einfach den Mauszeiger über einen Repository-Eintrag und verweilen Sie für einige Sekunden über dem Eintrag: Die Metadaten werden Ihnen in Form eines sogenannten Tooltips präsentiert. Anders als bei anderen Programmen, sind diese Hilfsinformationen jedoch deutlich mächtiger als gewohnt: Sie können einen solchen Tooltip mittels Druck auf die Taste F3 zu einem richtigen Dialog machen, den Sie beliebig verschieben und auch in der Größe ändern können. Außerdem sind diese RapidMiner Tooltips auch in der Lage, neben textuellen Informationen auch andere Elemente wie beispielsweise Tabellen mit den Metadaten aufzunehmen.

Beachten Sie bitte, dass die Metainformationen nicht zwingend sofort verfügbar sein müssen, sondern Sie das Einladen der Metadaten unter Umständen erst noch mit einem Klick auf einen Link innerhalb des Tooltips anstoßen müssen. Dieses Vorgehen verhindert, dass bei einem versehentlichen Ansehen der Tooltips der Repository-Einträge die unter Umständen doch recht großen Metadaten unmittelbar eingeladen werden müssen und RapidMiner auf diese Weise ausbremsen würden.

Tipp: Halten Sie den Mauszeiger kurz über einen Repository-Eintrag, um sich die Metadaten anzusehen oder erst einmal einzuladen. Handelt es sich bei dem Eintrag beispielsweise um ein Zwischenergebnis, können Sie leicht erkennen, welche Vorverarbeitung bereits stattgefunden hat.

Die folgende Abbildung zeigt, wie die Metadaten für den Golf-Datensatz aus dem mit RapidMiner mitgelieferten Beispielsverzeichnis aussehen. Zunächst erkennen Sie, dass der Datensatz 14 Beispiele („Number of examples“) und 5 Attribute enthält („Number of attributes“). Das Attribut mit dem Namen „Outlook“ ist nominal und nimmt die drei Werte „overcast“, „rain“ und „sunny“ an. Das Attribut „Temperature“ ist hingegen numerisch und nimmt Werte im Bereich von 64 bis 85 an – die Angabe ist natürlich in Fahrenheit. Das Attribut „Play“ schließlich ist wieder nominal, hat aber weiterhin eine spezielle Rolle: Es ist als „label“ markiert. Die Rolle ist kursiv gesetzt und steht noch vor dem Attributnamen.



Golf
Data Table
Number of examples = 14
5 attributes:

Role	Name	Type	Range	Missings	Comment
	Outlook	nominal	=[overcas...	= 0	
	Temperat...	integer	=[64 - 85]	= 0	
	Humidity	integer	=[65 - 96]	= 0	
	Wind	nominal	=[false, tr...	= 0	
label	Play	nominal	=[no, yes]	= 0	

Abbildung 5.6: Die Metadaten des Golfdatensatzes aus dem Beispielsverzeichnis des mit RapidMiner mitgelieferten Repositorys „Sample“. Sie finden den Datensatz namens „Golf“ im Verzeichnis „data“ in diesem Repository.

5.3.1 Metadatenpropagierung vom Repository durch den Prozess

Sie haben bereits gesehen, dass die oben beschriebenen Metadaten die eigentlichen Daten auf Ihrem weg durch den RapidMiner Prozess begleiten, bereits während Sie den Prozess erstellen. Wie schon zuvor erwähnt, ist es für diese Metadatenpropagierung und -transformation jedoch zwingend notwendig, dass Sie die Daten in einem RapidMiner Repository verwalten und die Metadaten von diesem erhalten können. Aus diesem Grund möchten wir noch einmal darauf hinweisen, dass die Verwendung des Repositorys zur Daten- und Prozessverwaltung für die Unterstützung während des Prozessdesigns erforderlich ist und hiermit noch einmal dringend empfohlen sei.

In diesem Abschnitt werden wir noch mal ein weiteres Beispiel für das Design eines Prozesses durchführen, wobei wir diesmal auf einen Datensatz aus dem RapidMiner Repository zurück greifen werden. Wir werden nun also erstmals den vollständigen Prozess vom Retrieval der Daten bis zur Erzeugung der Ergebnisse durchführen. Typischerweise würde diesem Prozess natürlich noch der Import der Daten in das Repository mittels einer der oben vorgestellten Methoden voran gehen, aber in diesem Fall verzichten wir auf diesen Schritt und verwenden stattdessen einfach einen der bereits von RapidMiner mitgelieferten Datensätze.

Laden Sie beispielsweise den mitgelieferten Datensatz Iris mit Hilfe eines Retrieve-Operators ein, indem Sie den betreffenden Eintrag (im gleichen Verzeichnis wie der bereits oben verwendete Golf-Datensatz) einfach in die Process View ziehen. Führen den Prozess aber noch nicht aus. Fügen Sie danach einen Normalize-Operator ein und verbinden Sie dessen Eingang mit dem Ausgang des Retrieve-Operators. Setzen sie den Parameter „method“ auf „range transformation“. Der Operator dient in dieser Einstellung dazu, numerische Werte neu zu skalieren, so dass das Minimum gerade 0 und das Maximum gerade 1 ist. Wählen Sie ein einzelnes Attribut aus, auf das Sie diese Transformation anwenden wollen, beispielsweise das Attribut „a3“. Setzen Sie dazu den Filtertyp „attribute filter type“ auf „single“ und wählen Sie das Attribut „a3“ am Parameter „attribute“ aus. Fahren Sie nun mit der Maus zunächst über den Ausgabeport von Retrieve und dann über den oberen Ausgangsport des Normalize-Operators. In beiden Fällen sehen Sie die Metadaten des Iris-Datensatzes. Sie werden jedoch bemerken, dass sich die Metadaten des gewählten Attributs verändert haben: Der Wertebereich von „a3“ ist nach der Transformation nun auf das Intervall [0,1] normalisiert. Oder präziser gesagt: Der Wertebereich von a3 *würde bei einer Ausführung* auf das Intervall [0,1] normalisiert werden.

Fügen Sie einen weiteren Operator ein, den Operatore „Discretize by Frequency“. Verbinden Sie diesen mit dem Normalize-Operator. Setzen Sie den Parameter „range name type“ auf „short“ und wählen Sie diesmal mit dem gleichen Mechanismus wie oben ein anderes Attribut aus, beispielsweise „a2“. Fahren Sie nun mit der Maus über den Ausgabeport des neuen Operators und beobachten Sie die Veränderung der Metadaten: Das ausgewählte Attribut ist nun nicht mehr numerisch sondern nominal und nimmt die Werte „range1“ und „range2“ an: Der Diskretisierungsoperator zerlegt den numerischen Wertebereich an einem Schwellwert und ersetzt Werte unterhalb dieses Wertes durch „range1“ und Werte oberhalb dieses Wertes durch „range2“. Der Schwellwert wird dabei automatisch

so gewählt, dass gleich viele Werte ober- und unterhalb liegen.

Wünschen Sie eine Unterteilung in mehrer als zwei Wertebereiche, passen Sie den Parameter „number of bins“ entsprechend an. Den Prozess und die angezeigten Metadaten sehen Sie in der folgenden Abbildung:

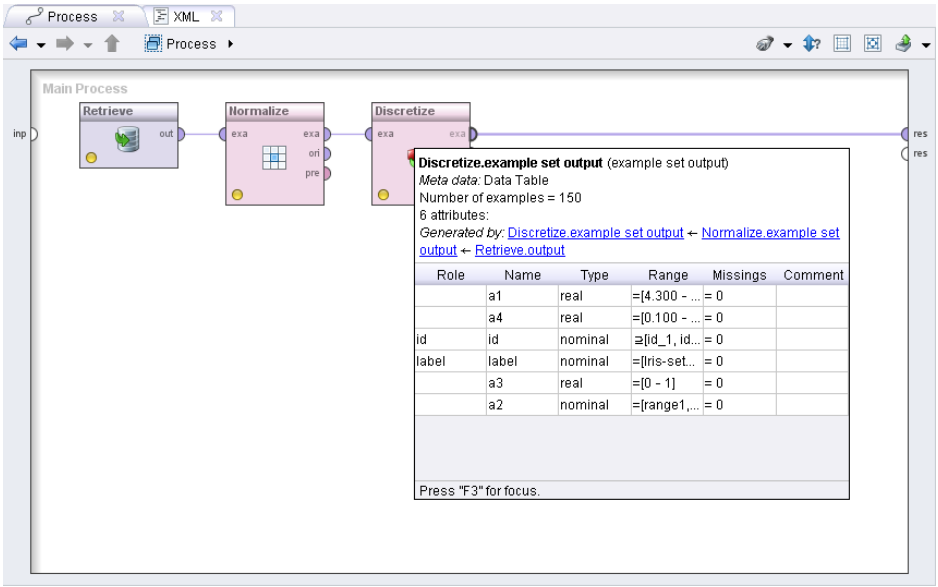

Process
XML
Process
Main Process
Retrieve
Normalize
Discretize
inp
out
exa
ori
pre
res
Discretize.example set output (example set output)
Meta data: Data Table
Number of examples = 150
6 attributes:
Generated by: Discretize.example set output ← Normalize.example set output ← Retrieve.output
Role | Name | Type | Range | Missings | Comment
a1 | real | =[4.300 - ... | = 0
a4 | real | =[0.100 - ... | = 0
id | id | nominal | ≥[id_1, id... | = 0
label | label | nominal | =[Iris-set... | = 0
a3 | real | =[0 - 1] | = 0
a2 | nominal | =[range1,... | = 0
Press "F3" for focus.

Abbildung 5.7: Metadatentransformation in RapidMiner.

Sie fragen sich sicher, warum der Parameter „range name type“ auf „short“ gesetzt werden musste. Probieren Sie es aus und setzen Sie ihn „long“. Führen Sie den Prozess aus, werden Sie sehen, dass die nominalen Werte nun ausdrucksstärker sind: Sie enthalten zusätzlich die Grenzen der erzeugten Intervalle. Dies ist praktisch, aber für den Prozess unerheblich. Die Informationen über die Intervallgrenzen sind jedoch nicht verfügbar, solange die Diskretisierung nicht tatsächlich durchgeführt wurde. Daher können sie für die Anzeige der Metadaten zur Entwicklungszeit des Prozesses nicht berücksichtigt werden. In den Metadaten ist dann für das diskretisierte Attribut der Wertebereich angegeben, dass es sich um die Obermenge der leeren Menge („{}“) handelt. Dies bedeutet, das die Metadaten nicht vollständig bekannt sind. In diesem Fall können wir also praktisch gar nichts über die erwarteten Metadaten sagen, eben außer, dass die Menge der nominalen Werte eine Obermenge der leeren Menge ist. Eine triviale Aussage, aber immerhin eine korrekte. Nicht in allen Fällen können die Metadaten zur Entwicklungszeit bereits vollständig ermittelt werden. Dies ist im Allgemeinen

immer dann der Fall, wenn die Metadaten wie hier von den tatsächlichen Daten abhängen. In diesem Fall versucht RapidMiner, so viel Information wie möglich über die Daten zu erhalten.

Rapid-I GmbH
Stockumer Str. 475
D-44227 Dortmund
Tel.: +49 (0) 231 425 786 90
E-Mail: contact@rapid-i.com
www.rapid-i.com